

Structural Markov graph laws for Bayesian model uncertainty

Simon Byrne and A. Philip Dawid

Abstract: This paper considers the problem of defining distributions over graphical structures. We propose an extension of the hyper Markov properties of Dawid and Lauritzen (1993), which we term *structural Markov properties*, for both undirected decomposable and directed acyclic graphs, which requires that the structure of distinct components of the graph be conditionally independent given the existence of a separating component. This allows the analysis and comparison of multiple graphical structures, while being able to take advantage of the common conditional independence constraints. Moreover, we show that these properties characterise exponential families, which form conjugate priors under sampling from compatible Markov distributions.

AMS 2000 subject classifications: Primary 62E10; secondary 62H10.

Keywords and phrases: Graphical models, structural estimation, hyper Markov laws, structural Markov laws.

1. Introduction

A graphical model consists of a graph and a probability distribution that satisfies a *Markov property* of the graph, being a set of conditional independence constraints encoded by the graph. Such models arise naturally in many statistical problems, such as contingency table analysis and covariance estimation.

Dawid and Lauritzen (1993) consider distributions over these distributions, which they term *laws* to emphasise the distinction from the underlying sampling distribution. Laws arise primarily in two contexts: as sampling distributions of estimators, and as prior and posterior distributions in Bayesian analyses. Specifically, Dawid and Lauritzen (1993) focus on hyper Markov laws, that exhibit conditional independence properties analogous to those of the distributions of the model. By exploiting such laws, it is possible to perform certain inferential tasks locally, for instance posterior laws can be calculated from subsets of the data pertaining to the parameters of interest.

Although other types of graphical model exist, we restrict ourselves to undirected decomposable graphs and directed acyclic graphs, which exhibit the special property that their Markov distributions can be constructed in a recursive fashion by taking *Markov combinations* of smaller components. In the case of undirected decomposable graphs, for any decomposition (A, B) of the graph \mathcal{G} , a Markov distribution is uniquely determined by the marginal distributions over A and B (Dawid and Lauritzen, 1993, Lemma 2.5). By a recursion argument, this is equivalent to specifying marginal distributions on cliques. A similar construction can be derived for directed acyclic graphs: the distribution of each

node on conditional on its parent set can be chosen arbitrarily, and the set of such distributions determines the joint distribution. As we demonstrate in section 5, this property can also be characterised in terms of a partitioning based on *ancestral sets*.

It is this partitioning that makes the notion of hyper Markov laws possible. In essence, these are laws for which the partitioned distributions exhibit conditional independence properties analogous to those of the underlying distributions. In the case of undirected decomposable graphs, a law \mathcal{L} for θ over $\mathfrak{P}(\mathcal{G})$, the set of Markov distributions with respect to \mathcal{G} , is *(weak) hyper Markov* if for any decomposition (A, B)

$$\tilde{\theta}_A \perp\!\!\!\perp \tilde{\theta}_B \mid \tilde{\theta}_{A \cap B} \quad [\mathcal{L}]. \quad (1.1)$$

Weak hyper Markov laws arise naturally as sampling distributions of maximum likelihood estimators of graphical models (Dawid and Lauritzen, 1993, Theorem 4.22). A more specific class of laws are those that satisfy the *strong hyper Markov property*, where for any decomposition (A, B)

$$\tilde{\theta}_{A|B} \perp\!\!\!\perp \tilde{\theta}_B \quad [\mathcal{L}]. \quad (1.2)$$

When used as prior laws in a Bayesian analysis, strong hyper Markov laws allow for local posterior updating, in that the posterior law of clique marginal distributions only depends on the data in the clique (Dawid and Lauritzen, 1993, Corollary 5.5).

Unfortunately, strong hyper Markov laws can be very restrictive. Dawid and Lauritzen (1993) outline two such laws: the hyper Dirichlet for contingency tables and the hyper inverse Wishart for Gaussian graphical models, which have Dirichlet and inverse Wishart marginal laws for the cliques.

The focus of this paper is to extend the hyper Markov concept to the structure of the graph itself. We study distributions over graphs, which we term *graph laws*, that exhibit similar conditional independence structure, termed *structural Markov properties*. These properties exhibit analogous local inference properties, and under minor assumptions, characterises a conjugate exponential family with strong hyper Markov laws.

The outline of the paper is as follows: in section 2 we introduce the notion of a semi-graphoid to define what we mean by structure. Section 3 develops the notion of a structural Markov property and characterise such laws for undirected decomposable graphs. Section 4 develops a similar notion for directed graphs consistent with a known ordering. In section 5 we consider the notion of Markov equivalence of directed graphs, and section 6 extends the structural Markov property to general equivalence classes of directed acyclic graphs. Finally, in section 7 we consider computational aspects of working with structural Markov graph laws.

1.1. Notation and definitions

Throughout the paper, unless otherwise noted, the set of vertices of the graph will be V . The edge set of a graph \mathcal{G} will be denoted by $\mathcal{E}(\mathcal{G})$: in the case of

undirected graphs, this will be sets of unordered pairs $\{u, v\}$, and in the directed case this will be a set of ordered pairs (u, v) , where $u, v \in V$. For any subset $A \subseteq V$, \mathcal{G}_A will denote the induced subgraph, with vertex set A and

Much of the notation will be borrowed from Dawid and Lauritzen (1993). Let $X = (X_v)_{v \in V}$ be a random vector, on some product space $\prod_{v \in V} \mathcal{X}_v$, with distribution denoted by P or θ . A family of distributions Θ for X will be termed a *model*, and a distribution over Θ will be termed a *law*, typically denoted by \mathcal{L} , with random variable θ .

For any subset $A \subseteq V$, X_A will denote the subvector $(X_v)_{v \in A}$, with P_A or θ_A denoting its marginal distribution. \mathcal{L}_A will denote the marginal law of θ_A . Furthermore, for any pair $A, B \subseteq V$, we can define $\theta_{A|B}$ to be the collection of conditional distributions of $X_A \mid X_B$ under θ , and $\mathcal{L}_{A|B}$ will be the marginal law of $\theta_{A|B}$. We will use \simeq to denote the existence of a bijective function, for instance we can write $(\theta_A, \theta_{V|A}) \simeq \theta$ for any $A \subseteq V$.

2. Semi-graphoid

When discussing the “structure” of graphical models, many authors refer to the graph itself, in particular, when discussing estimating the structure, they mean inferring the presence or absence of individual edges of the graph.

In this paper, we take the view that “structure” refers to the set of conditional independence statements, and that the graph is merely a representation of this structure. This distinction is an important one: it implies that graphs which encode the same set of conditional independence statements must be treated as identical, leading to the notion of *Markov equivalence*. A more subtle, but even more important point is that when investigating properties such as decompositions or ancestral sets we are, effectively, looking at properties of sets of conditional independencies.

To make this more concrete, we use the notion of a *semi-graphoid*, a special case of a *separoid* (Dawid, 2001), to describe the abstract properties of conditional independence.

Definition 2.1. Given a finite set V , a *semi-graphoid* is a set M of triples of the form $\langle A, B \mid C \rangle$, where $A, B, C \subseteq V$, satisfying the properties:

- S0 For all $A, B \subseteq V$, $\langle A, B \mid A \rangle \in M$,
- S1 If $\langle A, B \mid C \rangle \in M$, then $\langle B, A \mid C \rangle \in M$,
- S2 If $\langle A, B \mid C \rangle \in M$ and $D \subseteq A$, then $\langle D, B \mid C \rangle \in M$,
- S3 If $\langle A, B \mid C \rangle \in M$ and $D \subseteq A$, then $\langle A, B \mid C \cup D \rangle \in M$, and
- S4 If $\langle A, B \mid C \rangle \in M$ and $\langle A, D \mid B \cup C \rangle \in M$, then $\langle A, B \cup D \mid C \rangle \in M$.

These are defined so as to match the well-established properties of conditional independence (Dawid, 1979). Thus we can say a joint distribution P for $X = (X_v)_{v \in V}$ is *Markov* with respect to a semi-graphoid M if:

$$\langle A, B \mid C \rangle \in M \quad \Rightarrow \quad X_A \perp\!\!\!\perp X_B \mid X_C \quad [P].$$

We can define the semi-graphoid of a graph as the set of triples encoding its global Markov property: the semi-graphoid of an undirected graph \mathcal{G} is

$$\mathcal{M}(\mathcal{G}) = \{\langle A, B \mid C \rangle : A \text{ and } B \text{ are separated by } C \text{ in } \mathcal{G}\}, \quad (2.1)$$

and the semi-graphoid of a directed acyclic graph $\vec{\mathcal{G}}$ is the set

$$\mathcal{M}(\vec{\mathcal{G}}) = \{\langle A, B \mid C \rangle : A \text{ and } B \text{ are separated by } C \text{ in } \vec{\mathcal{G}}_{\text{an}(A \cup B \cup C)}^{\text{m}}\}. \quad (2.2)$$

That is, a distribution is Markov with respect to a graph if it is Markov with respect to the semi-graphoid of the graph.

Given a set S of such triples, we can define its *closure* \bar{S} to be the intersection of all semi-graphoids containing S . Conversely, we can define a subset $S \subseteq M$ to be *spanning set* if its closure is equal to M . Thus the local and pairwise Markov properties of undirected graphs, and the local and ordered Markov properties of directed acyclic graphs can be interpreted as spanning sets of the general semi-graphoid.

When we restrict ourselves to the case where \mathcal{G} is undirected decomposable graph, there is an even richer semi-graphoid structure. In particular, the set of decompositions is a spanning set of $\mathcal{M}(\mathcal{G})$ (Dawid and Lauritzen, 1993, Theorem 2.8).

We can even reduce the number of elements in the spanning set even further. Decomposable graphs admit a perfect sequence of cliques (also known as the running intersection property): for such a sequence C_1, C_2, \dots, C_k , define $H_i = C_1 \cup \dots \cup C_i$, and then the set

$$\{\langle C_i, H_{i-1} \mid C_i \cap H_{i-1} \rangle : i = 2, \dots, k\}$$

is a spanning set of the semi-graphoid.

Semi-graphoids have a natural projection operation: for any set $U \subseteq V$, we can define the projection onto U of a semi-graphoid M on V to be

$$M_U = \{\langle A, B \mid C \rangle \in M : A, B, C \subseteq U\}.$$

Under certain conditions, this can match the natural projection operation, the induced subgraph, of the underlying graph. For undirected graphs, Asmussen and Edwards (1983, Corollary 2.5) shows that $[\mathcal{M}(\mathcal{G})]_U = \mathcal{M}(G_U)$ if and only if \mathcal{G} is *collapsible* onto U . For directed acyclic graphs, we have the sufficient condition that if A is ancestral in $\vec{\mathcal{G}}$, then $[\mathcal{M}(\vec{\mathcal{G}})]_A = \mathcal{M}(\vec{\mathcal{G}}_A)$.

By utilising these properties, we can show that decompositions of undirected graphs can be thought of as decomposing the semi-graphoid.

Proposition 2.1. *Let (A, B) be a decomposition of an undirected graph \mathcal{G} . Then*

$$\mathcal{M}(\mathcal{G}_A) \cup \mathcal{M}(\mathcal{G}_B) \cup \{\langle A, B \mid A \cap B \rangle\}$$

is a spanning set of $\mathcal{M}(G)$.

Proof. This is a consequence of the fact that an undirected graph is uniquely specified by a decomposition, and the induced subgraph on each component of the decomposition. \square

3. Undirected structural Markov property

We now extend the hyper Markov framework to the case where the graph itself is a random variable $\tilde{\mathcal{G}}$. As the graph is a parameter of the model, we term its distribution a *graph law*, denoted by $\mathfrak{G}(\tilde{\mathcal{G}})$. Our aim is to identify and characterise hyper Markov-type properties for $\tilde{\mathcal{G}}$.

Let $\mathfrak{U}(A, B)$ denote the set of undirected decomposable graphs on V for which (A, B) is a decomposition. For a weak hyper Markov law $\mathcal{L}(\theta)$ on $\mathcal{G} \in \mathfrak{U}(A, B)$, recall that in (1.1), the support of θ_A is $\mathfrak{P}(\mathcal{G}_A)$, that is, \mathcal{G}_A is a parameter of X_A .

For a graph law $\mathfrak{G}(\tilde{\mathcal{G}})$ over $\mathfrak{U}(A, B)$, a straightforward way to extend the hyper Markov property in this case would be to require that

$$\tilde{\mathcal{G}}_A \perp\!\!\!\perp \tilde{\mathcal{G}}_B \mid \tilde{\mathcal{G}}_{A \cap B} \quad [\mathfrak{G}]. \quad (3.1)$$

Note that in this case the term $\tilde{\mathcal{G}}_{A \cap B}$ is redundant: if (A, B) is a decomposition of \mathcal{G} , then $\mathcal{G}_{A \cap B}$ must be complete, and so we are left with a statement of marginal independence $\tilde{\mathcal{G}}_A \perp\!\!\!\perp \tilde{\mathcal{G}}_B$.

A more general question remains: how might this property be extended to a graph law over all undirected graphs? A seemingly simple requirement is that (3.1) should hold whenever a decomposition exists. This motivates the following definition.

Definition 3.1 (Structural Markov property). A *covering pair* (of V) is any pair of sets (A, B) such that $A \cup B = V$. A graph law $\mathfrak{G}(\tilde{\mathcal{G}})$ over the set \mathfrak{U} of undirected decomposable graphs on V is *structurally Markov* if for any covering pair (A, B) , $\tilde{\mathcal{G}}_A$ is independent of $\tilde{\mathcal{G}}_B$, conditional on (A, B) being a decomposition of $\tilde{\mathcal{G}}$. In other words, that

$$\tilde{\mathcal{G}}_A \perp\!\!\!\perp \tilde{\mathcal{G}}_B \mid \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\} \quad [\mathfrak{G}], \quad (3.2)$$

where $\mathfrak{U}(A, B)$ is the set of decomposable graphs for which (A, B) is a decomposition.

In essence, the structural Markov property states that the structure of different induced subgraphs are conditionally independent given that they are in separate parts of a decomposition. See Figure 1 for a depiction.

The use of braces on the right-hand side of (3.2) is to emphasise that the conditional independence is defined with respect to the *event* $\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)$, and not a random variable as in the Markov and hyper Markov properties. In other words, we do not assume $\tilde{\mathcal{G}}_A \perp\!\!\!\perp \tilde{\mathcal{G}}_B \mid \tilde{\mathcal{G}} \notin \mathfrak{U}(A, B)$.

3.1. Products and projections

Lemmas 2.5 and 3.3 of Dawid and Lauritzen (1993) allow the construction of Markov distributions and hyper Markov laws in a piecewise manner over the cliques, via a conditional product operation. The same arguments can also be applied to graphs laws.

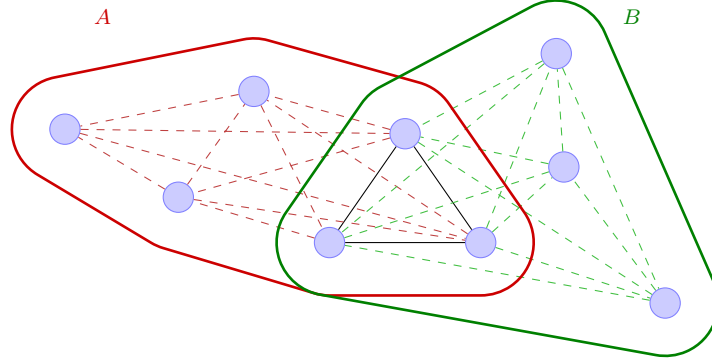


Fig 1: A representation of the structural Markov property for undirected graphs. Conditional on (A, B) being a decomposition, the existence of the remaining edges in $\tilde{\mathcal{G}}_A$ (---) is independent of those in $\tilde{\mathcal{G}}_B$ (---).

Proposition 3.1. *Let \mathcal{H} and \mathcal{J} be two graphs on A and B respectively, such that both $\mathcal{H}_{A \cap B}$ and $\mathcal{J}_{A \cap B}$ are complete. Then there exists a unique graph \mathcal{G} on $A \cup B$ such that*

- (i) $\mathcal{G}_A = \mathcal{H}$,
- (ii) $\mathcal{G}_B = \mathcal{J}$, and
- (iii) $A \cap B$ separates A and B in \mathcal{G} .

Proof. To satisfy (i) and (ii), the edge set must contain $\mathcal{E}(\mathcal{H}) \cup \mathcal{E}(\mathcal{J})$. It cannot contain any additional edges $\{u, v\}$, as this would violate either (i), if $\{u, v\} \subseteq A$; (ii), if $\{u, v\} \subseteq B$, or (iii), if $u \in A \setminus B$ and $v \in B \setminus A$. \square

We define the resulting graph to be the *graph product*, denoted by

$$\mathcal{G} = \mathcal{H} \otimes \mathcal{J},$$

and the completeness requirement on the intersection $A \cap B$ implies that (A, B) will be a decomposition of \mathcal{G} . The graph product provides a very useful characterisation of the structural Markov property.

Proposition 3.2. *A graph law \mathfrak{G} is structurally Markov if and only if for every covering pair (A, B) , and every $\mathcal{G}, \mathcal{G}' \in \mathfrak{U}(A, B)$,*

$$\pi(\mathcal{G}) \pi(\mathcal{G}') = \pi(\mathcal{G}_A \otimes \mathcal{G}'_B) \pi(\mathcal{G}'_A \otimes \mathcal{G}_B) \quad (3.3)$$

where π is the density of \mathfrak{G} with respect to the counting measure on \mathfrak{U} .

Proof. Note that both $\mathcal{G}_A \otimes \mathcal{G}'_B, \mathcal{G}'_A \otimes \mathcal{G}_B \in \mathfrak{U}(A, B)$. Furthermore, the conditional density of a structural Markov law is of the form

$$\pi(\mathcal{G} \mid \{\mathcal{G} \in \mathfrak{U}(A, B)\}) = \pi(\mathcal{G}_A \mid \mathfrak{U}(A, B)) \pi(\mathcal{G}_B \mid \mathfrak{U}(A, B)).$$

The result follows by substitution into (3.3). \square

The structural Markov property has an inherent divisibility property that arises on subgraphs induced by decompositions. First we require the following lemma.

Lemma 3.3. *Let (A, B) be a decomposition of a graph \mathcal{G} , and (S, T) a covering pair of A with $A \cap B \subseteq T$. Then (S, T) is a decomposition of \mathcal{G}_A if and only if $(S, T \cup B)$ is a decomposition of \mathcal{G} .*

Proof. Recall that W separates U and V in \mathcal{G} if and only if $\langle U, V \mid W \rangle \in \mathcal{M}(\mathcal{G})$. Since (S, T) is a covering pair of A , $\langle S \cup T, B \mid S \cap B \rangle \in \mathcal{M}(\mathcal{G})$, and hence $\langle S, B \mid T \rangle \in \mathcal{M}(\mathcal{G})$. If (S, T) is a decomposition of \mathcal{G}_A , then $\langle S, T \mid S \cap T \rangle \in \mathcal{M}(\mathcal{G}_A)$, which implies that $\langle S, B \cup T \mid T \cap S \rangle \in \mathcal{M}(\mathcal{G})$. Since $\mathcal{G}_{(S \cup B) \cap T} = \mathcal{G}_{T \cap S}$ is complete, $(S \cup B, T)$ is a decomposition of \mathcal{G} .

The converse result is follows by the reverse argument. \square

Theorem 3.4. *Let $\mathfrak{G}(\tilde{\mathcal{G}})$ be a structurally Markov graph law: then the conditional law for $\tilde{\mathcal{G}}_A \mid \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\}$ is also structurally Markov.*

Proof. Let (S, T) be a covering pair of A : If we restrict $\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)$, then $\tilde{\mathcal{G}}_{A \cap B}$ must be complete. As we are only interested in the case where (S, T) is a decomposition of $\tilde{\mathcal{G}}_A$, then $A \cap B$ must be a subset of either S or T : without loss of generality, we may assume $A \cap B \subseteq T$.

$(S, T \cup B)$ is a covering pair of V , so by the structural Markov property:

$$\tilde{\mathcal{G}}_S \perp\!\!\!\perp \tilde{\mathcal{G}}_{T \cup B} \mid \{\tilde{\mathcal{G}} \in \mathfrak{U}(S, T \cup B)\}.$$

If $\mathbb{1}_E$ is the indicator variable of an event E , we may can write:

$$\tilde{\mathcal{G}}_S \perp\!\!\!\perp (\tilde{\mathcal{G}}_T, \mathbb{1}_{\tilde{\mathcal{G}}_{T \cup B} \in \mathfrak{U}(T, B)}) \mid \{\tilde{\mathcal{G}} \in \mathfrak{U}(S, T \cup B)\}.$$

By the properties of conditional independence (Dawid, 1979), the term $\mathbb{1}_{\tilde{\mathcal{G}}_{T \cup B} \in \mathfrak{U}(T, B)}$ may be moved to the right-hand side. Furthermore, we are only interested in the case where it equals 1, hence we can write:

$$\tilde{\mathcal{G}}_S \perp\!\!\!\perp \tilde{\mathcal{G}}_T \mid \{\mathcal{G}_{T \cup B} \in \mathfrak{U}(T, B)\}, \{\tilde{\mathcal{G}} \in \mathfrak{U}(S, T \cup B)\}.$$

By Lemma 3.3, $\tilde{\mathcal{G}}_{T \cup B} \in \mathfrak{U}(T, B)$ if and only if $\tilde{\mathcal{G}} \in \mathfrak{U}(S \cup T, B) = \mathfrak{U}(A, B)$.

$$\tilde{\mathcal{G}}_S \perp\!\!\!\perp \tilde{\mathcal{G}}_T \mid \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\}, \{\tilde{\mathcal{G}} \in \mathfrak{U}(S, T \cup B)\}.$$

Again, by Lemma 3.3, $\tilde{\mathcal{G}} \in \mathfrak{U}(S, T \cup B)$ if and only if $\tilde{\mathcal{G}}_A \in \mathfrak{U}(S, T)$, hence:

$$\tilde{\mathcal{G}}_S \perp\!\!\!\perp \tilde{\mathcal{G}}_T \mid \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\}, \{\tilde{\mathcal{G}}_A \in \mathfrak{U}(S, T)\}. \quad \square$$

3.2. Structural meta Markov property

Dawid and Lauritzen (1993) define *meta Markov models*: a set of probability Markov property distributions that exhibit *conditional variation independence*, denoted by the ternary relation $(\cdot \ddagger \cdot \mid \cdot)$, in place of the conditional probabilistic independence of hyper Markov laws. Analogous properties can be defined in the structural context.

Definition 3.2 (Structural meta Markov property). For a family of undirected decomposable graphs \mathfrak{F} and a covering pair (A, B) , let $\mathfrak{F}(A, B) = \mathfrak{F} \cap \mathfrak{U}(A, B)$. Then \mathfrak{F} is *structurally meta Markov* if for every covering pair (A, B) ,

$$\mathcal{G}_A \nmid \mathcal{G}_B \mid \{\mathcal{G} \in \mathfrak{F}(A, B)\}.$$

As with probabilistic independence, we can characterise it in terms of the graph product operation.

Theorem 3.5. *A family of undirected decomposable graphs \mathfrak{F} is structurally meta Markov if and only if $\mathcal{G}_A \otimes \mathcal{G}'_B \in \mathfrak{F}$ for all $\mathcal{G}, \mathcal{G}' \in \mathfrak{F}(A, B)$.*

Proof. This follows directly from Proposition 3.1. \square

Theorem 3.5 is particularly useful in that if a family of graphs is characterised by a specific property, we can show that it is structurally meta Markov if this property is preserved under the graph product operation.

Example 3.1. The set of undirected decomposable graphs whose clique size is bounded above by some integer n is structurally meta Markov. To see this, note that a clique of $\mathcal{G}_A \otimes \mathcal{G}'_B$ must be a clique of either \mathcal{G}_A or \mathcal{G}'_B (and hence of either \mathcal{G} or \mathcal{G}'). Therefore, the graph product operation cannot increase the size of the largest clique.

An interesting special case is $n = 2$, which is the set of forests on V .

Example 3.2. Consider two graphs $\mathcal{G}^L, \mathcal{G}^U \in \mathfrak{U}$ such that $\mathcal{E}(\mathcal{G}^L) \subseteq \mathcal{E}(\mathcal{G}^U)$. Then the “sandwich” set between the two graphs,

$$\{\mathcal{G} \in \mathfrak{U} : \mathcal{E}(\mathcal{G}^L) \subseteq \mathcal{E}(\mathcal{G}) \subseteq \mathcal{E}(\mathcal{G}^U)\},$$

is structurally meta Markov. This follows from the fact that an edge can only appear in a graph product if it is in one of the elements of the product.

As with hyper Markov laws, a structural meta Markov family is a necessary condition for the existence of a structural Markov law.

Theorem 3.6. *The support of a structurally Markov graph law is a structurally meta Markov family.*

Proof. Let \mathfrak{F} be the support of the structurally Markov graph law \mathfrak{G} with density π . By Proposition 3.2, if $\mathcal{G}, \mathcal{G}' \in \mathfrak{F}(A, B)$ and both $\pi(\mathcal{G})$ and $\pi(\mathcal{G}')$ are non-zero, then $\pi(\mathcal{G}_A \otimes \mathcal{G}'_B)$ must also be non-zero, and hence in $\mathfrak{F}(A, B)$. Therefore, by Theorem 3.5, \mathfrak{F} is structurally meta Markov. \square

3.3. Compatible distributions and laws

We now investigate how the structural Markov property interacts with the Markov and hyper Markov properties. In order to do this, we need to define families of distributions and laws for every graph.

Definition 3.3. Let $X = (X_v)_{v \in V}$ be a random variable, $\mathfrak{F} \subseteq \mathfrak{U}$, and $\vartheta = \{\theta^{(\mathcal{G})} : \mathcal{G} \in \mathfrak{F}\}$ be a family of probability distributions for X . We write $X \sim \vartheta \mid \tilde{\mathcal{G}}$ if, given $\tilde{\mathcal{G}} = \mathcal{G}$, $X \sim \theta^{(\mathcal{G})}$. Then ϑ is *compatible* if

- (i) for each $\mathcal{G} \in \mathfrak{U}$, X is Markov with respect to \mathcal{G} under $\theta^{(\mathcal{G})}$, and
- (ii) $\theta_C^{(\mathcal{G})} = \theta_C^{(\mathcal{G}')}$ whenever $C \subseteq V$ induces a complete subgraph in both $\mathcal{G}, \mathcal{G}' \in \mathfrak{U}$.

Likewise, let $\mathfrak{L} = \{\mathcal{L}^{(\mathcal{G})} : \mathcal{G} \in \mathfrak{F}\}$ be a family of laws for the parameters $\tilde{\theta}$ of a family of distributions on X . Again, we can write $\tilde{\theta} \sim \mathfrak{L} \mid \tilde{\mathcal{G}}$ if, given $\tilde{\mathcal{G}} = \mathcal{G}$, $\tilde{\theta} \sim \mathcal{L}^{(\mathcal{G})}$. Then \mathfrak{L} is *hyper compatible* if

- (i) for all $\mathcal{G} \in \mathfrak{U}$, $\mathcal{L}^{(\mathcal{G})}$ is a weak hyper Markov law on \mathcal{G} , and
- (ii) $\mathcal{L}_C^{(\mathcal{G})} = \mathcal{L}_C^{(\mathcal{G}')}$ if C induces a complete subgraph in both $\mathcal{G}, \mathcal{G}' \in \mathfrak{U}$.

Remark. Dawid and Lauritzen (1993, section 6.2) originally used the term compatible to refer to what we term the hyper compatible case: we introduce the distinction so as to extend the terminology to the distributional (non-hyper) case.

When they satisfy the compatibility condition, both ϑ and \mathfrak{L} are characterised entirely by $\theta^{(\mathcal{G}^{(V)})}$ and $\mathcal{L}^{(\mathcal{G}^{(V)})}$ respectively, where $\mathcal{G}^{(V)}$ is the complete graph on V .

A law induces a marginal distribution, referred to as the *predictive distribution* in Bayesian problems. Therefore a family of laws will also induce a family of distributions. Although in general hyper compatibility will not imply compatibility, there is one important special case.

Proposition 3.7. *Let \mathfrak{L} be a family of laws such that each law $\mathcal{L}^{(\mathcal{G})} \in \mathfrak{L}$ is strong hyper Markov. Then the marginal family of distributions is hyper compatible.*

Proof. By Dawid and Lauritzen (1993, Proposition 5.6), the marginal distribution of a strong hyper Markov law is Markov with respect to the same graph. The result follows by noting that the marginal distribution on a complete subgraph is a function of the marginal law. \square

A graph law $\mathfrak{G}(\tilde{\mathcal{G}})$ combined with a compatible set of distributions ϑ defines a joint distribution $(\mathfrak{G}, \vartheta)$ for $(\tilde{\mathcal{G}}, X)$ under which $X \mid \tilde{\mathcal{G}} = \mathcal{G} \sim \theta^{(\mathcal{G})}$. Likewise, \mathfrak{G} combined with a set of hyper compatible laws \mathfrak{L} defines a joint law $(\mathfrak{G}, \mathfrak{L})$ for $(\tilde{\mathcal{G}}, \tilde{\theta})$, and so a joint distribution on $(\tilde{\mathcal{G}}, \tilde{\theta}, X)$.

The key conditional independence property of any such joint distribution or law can be characterised as follows.

Proposition 3.8. *If $\tilde{\mathcal{G}}$ has a graph law \mathfrak{G} , and $X \sim \vartheta$ for a compatible family ϑ , then*

$$X_A \perp\!\!\!\perp \tilde{\mathcal{G}}_B \mid \tilde{\mathcal{G}}_A, \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\} \quad [\mathfrak{G}, \vartheta].$$

Similarly, if $\tilde{\mathcal{G}}$ has a graph law \mathfrak{G} , and $\tilde{\theta} \sim \mathfrak{L}$ for a hyper compatible family \mathfrak{L} , then

$$\tilde{\theta}_A \perp\!\!\!\perp \tilde{\mathcal{G}}_B \mid \tilde{\mathcal{G}}_A, \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\} \quad [\mathfrak{G}, \mathfrak{L}].$$

Proof. If $\mathcal{G} \in \mathfrak{U}(A, B)$, then \mathcal{G}_A is uniquely determined by its cliques. Therefore the distribution of X_A and law of $\tilde{\theta}_A$ are each fixed. \square

When combined with the structural Markov property, we obtain some useful results.

Theorem 3.9. *If $\tilde{\mathcal{G}}$ has a structurally Markov graph law \mathfrak{G} , and X has a distribution from a compatible set ϑ , then*

$$(X_A, \tilde{\mathcal{G}}_A) \perp\!\!\!\perp (X_B, \tilde{\mathcal{G}}_B) \mid X_{A \cap B}, \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\} \quad [\mathfrak{G}, \vartheta].$$

Proof. The Markov property states that under $[\mathfrak{G}, \vartheta]$,

$$X_A \perp\!\!\!\perp X_B \mid X_{A \cap B}, \tilde{\mathcal{G}}, \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\}. \quad (3.4)$$

Since if $\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)$, then $\tilde{\mathcal{G}} \simeq (\tilde{\mathcal{G}}_A, \tilde{\mathcal{G}}_B)$, we can rewrite (3.4) as

$$X_A \perp\!\!\!\perp X_B \mid X_{A \cap B}, \tilde{\mathcal{G}}_A, \tilde{\mathcal{G}}_B, \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\}. \quad (3.5)$$

As a consequence of Proposition 3.8,

$$X_A \perp\!\!\!\perp \tilde{\mathcal{G}}_B \mid X_{A \cap B}, \tilde{\mathcal{G}}_A, \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\}, \quad (3.6)$$

and combined with (3.5),

$$X_A \perp\!\!\!\perp (X_B, \tilde{\mathcal{G}}_B) \mid X_{A \cap B}, \tilde{\mathcal{G}}_A, \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\}. \quad (3.7)$$

Furthermore, by the structural Markov property and Proposition 3.8,

$$\tilde{\mathcal{G}}_A \perp\!\!\!\perp (X_B, \tilde{\mathcal{G}}_B) \mid \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\}, \quad (3.8)$$

and we can further condition on $X_{A \cap B}$. The result follows from this and (3.7). \square

Corollary 3.10. *If $\tilde{\mathcal{G}}$ has a structurally Markov graph law, and X has a distribution from a compatible set ϑ , then the posterior graph law for $\tilde{\mathcal{G}}$ is structurally Markov.*

Proof. By Theorem 3.9 and the axioms of conditional independence, we easily obtain

$$\tilde{\mathcal{G}}_A \perp\!\!\!\perp \tilde{\mathcal{G}}_B \mid X, \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\}. \quad \square$$

We can also apply similar arguments at the hyper level.

Theorem 3.11. *If $\tilde{\mathcal{G}}$ has a structurally Markov graph law \mathfrak{G} , and θ has a law from a hyper compatible set \mathfrak{L} , then*

$$(\theta_A, \tilde{\mathcal{G}}_A) \perp\!\!\!\perp (\theta_B, \tilde{\mathcal{G}}_B) \mid \theta_{A \cap B}, \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\} \quad [\mathfrak{G}, \mathfrak{L}].$$

Furthermore, if each law $\mathcal{L}^{(\mathcal{G})} \in \mathfrak{L}$ is strong hyper Markov with respect to \mathcal{G} , then

$$(\theta_A, \tilde{\mathcal{G}}_A) \perp\!\!\!\perp (\theta_{B|A}, \tilde{\mathcal{G}}_B) \mid \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\} \quad [\mathfrak{G}, \mathfrak{L}].$$

Proof. The proof for the first case is the same as in Theorem 3.9. The proof for the strong case follows similar steps, except starting with the strong hyper Markov property:

$$\theta_A \perp\!\!\!\perp \theta_{B|A} \mid \tilde{\mathcal{G}}, \{\tilde{\mathcal{G}} \in \mathfrak{U}(A, B)\}. \quad \square$$

Hyper compatible sets of strong hyper Markov laws have the additional advantage that the posterior graph law will also be structurally Markov: this follows from Theorem 3.9 and Dawid and Lauritzen (1993, Proposition 5.6), which states that the marginal distribution of the data under a strong hyper Markov law is Markov. Furthermore, the posterior family of graph laws $\{\mathcal{L}^{(\mathcal{G})}(\cdot \mid X) : \mathcal{G} \in \mathfrak{U}\}$ will maintain hyper compatibility.

3.4. Clique vector

We show that the family of structural Markov laws forms an exponential family of conjugate distributions for Bayesian updating under compatible sampling.

Definition 3.4. Define the *completeness vector* of a graph to be the function $c : \mathfrak{U} \rightarrow \{0, 1\}^{2^V}$, such that for each $A \subseteq V$,

$$c_A(\mathcal{G}) = \begin{cases} 1 & \text{if } \mathcal{G}_A \text{ is complete,} \\ 0 & \text{otherwise.} \end{cases}$$

Furthermore, define the *clique vector* of a graph $t : \mathfrak{U} \rightarrow \mathbb{Z}^{2^V}$ to be the Möbius inverse of c by *superset* inclusion:

$$t_B(\mathcal{G}) = \sum_{A \supseteq B} (-1)^{|A \setminus B|} c_A(\mathcal{G}). \quad (3.9)$$

In the language of Studený (2005b), c and t are both *imsets*

The decomposition of c and t mirrors that of the graph.

Lemma 3.12. *If $\mathcal{G} \in \mathfrak{U}(A, B)$, then*

$$c(\mathcal{G}) = [c(\mathcal{G}_A)]^0 + [c(\mathcal{G}_B)]^0 - [c(\mathcal{G}_{A \cap B})]^0, \quad \text{and} \quad (3.10)$$

$$t(\mathcal{G}) = [t(\mathcal{G}_A)]^0 + [t(\mathcal{G}_B)]^0 - [t(\mathcal{G}_{A \cap B})]^0; \quad (3.11)$$

where $[\cdot]^0$ denotes the expansion of a vector with zeroes to the required coordinates.

Proof. A subset $U \subseteq V$ induces a complete subgraph of $\mathcal{G} \in \mathfrak{U}(A, B)$ if and only if it induces a complete subgraph from \mathcal{G}_A , \mathcal{G}_B or both. (3.10) follows by the inclusion-exclusion principle. (3.11) may then be obtained by substitution into (3.9). \square

Theorem 3.13. *For any decomposable graph $\mathcal{G} \in \mathfrak{U}$ and $A \subseteq V$,*

$$t_A(\mathcal{G}) = \begin{cases} 1 & \text{if } A \in \text{cl}(\mathcal{G}), \\ -\nu_{\mathcal{G}}(A) & \text{if } A \in \text{sep}(\mathcal{G}), \text{ and} \\ 0 & \text{otherwise;} \end{cases}$$

where $\text{cl}(\mathcal{G})$ are the cliques of \mathcal{G} , and $\text{sep}(\mathcal{G})$ are the clique separators, and each separator S has multiplicity $\nu_{\mathcal{G}}(S)$.

Proof. For any $C \subseteq V$, let $\mathcal{G}^{(C)}$ be the graph on V whose edges are the set of all pairs $\{u, v\} \subseteq C$ (that is, complete on C and empty elsewhere). Then it is straightforward to see that

$$t_A(\mathcal{G}_C^{(C)}) = \begin{cases} 1 & \text{if } A = C, \\ 0 & \text{otherwise.} \end{cases}$$

Now let C_1, \dots, C_k be a perfect ordering of the cliques of G , and S_2, \dots, S_k be the corresponding separators. By Lemma 3.12, it follows that

$$t(\mathcal{G}) = \sum_{i=1}^k t(\mathcal{G}_{C_i}^{(C_i)}) - \sum_{i=2}^k t(\mathcal{G}_{S_i}^{(S_i)}). \quad \square$$

Objects similar to the clique vector have arisen in several contexts. Notably, it appears to be equivalent to the index v of Lauritzen, Speed and Vijayan (1984, Definition 5), which is characterised in a combinatorial manner. It is also closely related to the *standard imset* of Studený (2005b), which is equal to

$$t(\mathcal{G}^{(V)}) - t(\mathcal{G})$$

where $\mathcal{G}^{(V)}$ is the complete graph.

The algorithm of Wormald (1985) for the enumeration of decomposable graphs is based on a generating function for the vector $\mathbb{R}^{|V|}$ that he termed the “maximal clique vector”, and is equivalent to

$$\text{mcv}_k(\mathcal{G}) = \sum_{A \subseteq V: |A|=k} t_A(\mathcal{G}), \quad k = 1, \dots, |V|$$

Proposition 3.14. *For any $\mathcal{G} \in \mathfrak{U}$, the vector $t(\mathcal{G})$ has the following properties:*

(i)

$$\sum_{A \subseteq V} t_A(\mathcal{G}) = 1,$$

(ii) *for each $v \in V$*

$$\sum_{A \ni v} t_A(\mathcal{G}) = 1,$$

(iii)

$$\sum_{A \subseteq V} |A| t_A(\mathcal{G}) = |V|, \text{ and}$$

(iv)

$$\sum_{A \subseteq V} \binom{|A|}{2} t_A(\mathcal{G}) = |\mathcal{E}(\mathcal{G})|.$$

Proof. These all follow from Theorem 3.13 and the inclusion-exclusion principle. \square

3.5. Clique exponential family

Definition 3.5. The *clique exponential family* is the exponential family of graph laws over $\mathfrak{F} \subseteq \mathfrak{U}$, with t as a natural statistic (with respect to the uniform measure on \mathfrak{U}). That is, laws in the family have densities of the form

$$\pi_\omega(\mathcal{G}) = \frac{1}{Z(\omega)} \exp\{\omega \cdot t(\mathcal{G})\}, \quad \mathcal{G} \in \mathfrak{F}, \quad \omega \in \mathbb{R}^{2^V},$$

where $Z(\omega)$ is the normalisation constant, which will generally be hard to compute.

Equivalently, the distribution can be parameterised in terms of c ,

$$\pi_\omega(\mathcal{G}) = \frac{1}{Z(\omega)} \exp \left\{ \left(\sum_{B \subseteq A} (-1)^{|A \setminus B|} \omega_A \right)_{A \subseteq V} \cdot c(\mathcal{G}) \right\},$$

but t is more useful due to the fact that it is sparse (by Theorem 3.13) and, as we shall see, is the natural statistic for posterior updating.

Note that this distribution is over-parametrised: by Proposition 3.14 (i) and (ii), there are $|V| + 1$ linear relationships in $t(\mathcal{G})$. For the purpose of identifiability, we could define a normalised vector ω^* as

$$\omega_A^* = \omega_A + (|A| - 1)\omega_\emptyset - \sum_{v \in A} \omega_{\{v\}},$$

such that $\pi_\omega = \pi_{\omega^*}$, and $\omega_{\{v\}}^* = \omega_\emptyset^* = 0$ for all $v \in V$.

Theorem 3.15. Let \mathfrak{G} be a graph law whose support is \mathfrak{U} . Then \mathfrak{G} is structurally Markov if and only if it is a member of the clique exponential family.

Proof. For any $C \subseteq V$, define $\mathcal{G}^{(C)}$ as in the proof of Theorem 3.13, and let \mathfrak{G} have density π .

Suppose that \mathfrak{G} is structurally Markov. For any $\mathcal{G} \in \mathfrak{U}$, let C_1, \dots, C_k be a perfect ordering of the cliques, and let S_2, \dots, S_k be the corresponding separators, and $H_i = C_1 \cup \dots \cup C_i$. Furthermore, recursively define the graphs

$$\mathcal{G}^{*(j)} = \begin{cases} \mathcal{G}^{(C_1)} & \text{if } j = 1, \\ \mathcal{G}_{H_{j-1}}^{*(j-1)} \otimes \mathcal{G}_{(V \setminus H_{j-1}) \cup S_j}^{(C_j)} & \text{if } j = 2, \dots, k. \end{cases}$$

By Proposition 3.2, for each $j = 2, \dots, k$

$$\pi(\mathcal{G}^{*(j)})\pi(\mathcal{G}^{(S_j)}) = \pi(\mathcal{G}^{*(j-1)})\pi(\mathcal{G}^{(C_j)})$$

Note that $\mathcal{G}^{*(k)} = \mathcal{G}$, then by induction,

$$\pi(\mathcal{G}) = \frac{\prod_{j=1}^k \pi(\mathcal{G}^{(C_j)})}{\prod_{j=2}^k \pi(\mathcal{G}^{(S_j)})} \propto \exp\{\omega \cdot t(\mathcal{G})\},$$

by Theorem 3.13, where $\omega_C = \log \pi(\mathcal{G}^{(C)})$.

To show the converse let $(\omega)_A = (\omega_S)_{S \subseteq A}$. By Lemma 3.12,

$$\begin{aligned} \pi(\mathcal{G}_A | \mathcal{G}_B, \{\mathcal{G} \in \mathfrak{U}(A, B)\}) &\propto \exp \{(\omega)_A \cdot t(\mathcal{G}_A) + (\omega)_B \cdot t(\mathcal{G}_B) - (\omega)_{A \cap B} \cdot t(\mathcal{G}_{A \cap B})\} \\ &\propto \exp \{(\omega)_A \cdot t(\mathcal{G}_A) - (\omega)_{A \cap B} \cdot t(\mathcal{G}_{A \cap B})\} \\ &\propto \pi(\mathcal{G}_A | \{\mathcal{G} \in \mathfrak{U}(A, B)\}). \square \end{aligned}$$

Remark. It is possible to weaken the condition of full support, for example the same argument applies to any family \mathfrak{F} with the property that if $\mathcal{G} \in \mathfrak{F}$ and C is a clique of \mathcal{G} , then $\mathcal{G}^{(C)} \in \mathfrak{F}$. This includes Example 3.1, and Example 3.2 if \mathcal{G}^L is the sparse graph.

A very similar family was proposed by Bornn and Caron (2011), however their family allows the use of different parameters for cliques and separators, which will generally not be structurally Markov.

Example 3.3 (Giudici and Green, 1999; Brooks, Giudici and Roberts, 2003, section 8). The simplest example of such a distribution is the uniform distribution over \mathfrak{U} , corresponding to $\omega = 0$.

Example 3.4 (Madigan and Raftery, 1994; Jones et al., 2005). Another common approach is to use a set of $\binom{|V|}{2}$ independent Bernoulli variables with probability ψ to indicate edge inclusion (*i.e.* an Erdős–Rényi random graph), conditional on $\tilde{\mathcal{G}}$ being decomposable. The density of such a law is of the form

$$\pi(\mathcal{G}) \propto \psi^{|\mathcal{E}(\mathcal{G})|} (1 - \psi)^{\binom{p}{2} - |\mathcal{E}(\mathcal{G})|} \propto \left(\frac{\psi}{1 - \psi} \right)^{|\mathcal{E}(\mathcal{G})|}.$$

By Proposition 3.14 (iv), it follows that this distribution has

$$\omega_A = \binom{|A|}{2} \log \left(\frac{\psi}{1 - \psi} \right).$$

More generally, if each edge e has its own probability ψ_e , then

$$\omega_A = \sum_{e \in \binom{A}{2}} \log \left(\frac{\psi_e}{1 - \psi_e} \right).$$

Example 3.5 (Armstrong et al., 2009). For comparison, it is useful to consider a non-structurally Markov graph law. Define the distribution over the number of edges to be uniform, and the conditional distribution over the set of graphs with a fixed number of edges to be uniform. This has density of the form

$$\pi(\mathcal{G}) = \frac{1}{\binom{p}{2} + 1} \frac{1}{|\{\mathcal{G}' \in \mathfrak{U} : |\mathcal{E}(\mathcal{G}')| = |\mathcal{E}(\mathcal{G})|\}|}.$$

Specifically, for the case $V = \{1, 2, 3\}$, then:

$$\begin{aligned}\pi\left(\begin{array}{ccc} \textcircled{1} & \textcircled{2} & \textcircled{3} \end{array}\right) &= \frac{1}{4} \\ \pi\left(\begin{array}{ccc} \textcircled{1} & \text{---} \textcircled{2} & \textcircled{3} \end{array}\right) &= \frac{1}{12} \\ \pi\left(\begin{array}{ccc} \textcircled{1} & \textcircled{2} & \text{---} \textcircled{3} \end{array}\right) &= \frac{1}{12} \\ \pi\left(\begin{array}{ccc} \textcircled{1} & \text{---} \textcircled{2} & \text{---} \textcircled{3} \end{array}\right) &= \frac{1}{12}\end{aligned}$$

From this it follows that $\tilde{\mathcal{G}}_{\{1,2\}} \not\sim \tilde{\mathcal{G}}_{\{2,3\}} \mid \{\tilde{\mathcal{G}} \in \mathfrak{U}(\{1,2\}, \{2,3\})\}$, and hence the law cannot be structurally Markov.

3.6. Posterior updating

We saw in Corollary 3.10 that if the sampling distributions are compatible, then posterior updating will preserve the structural Markov property. In this section we show that this updating may be performed locally, with the exponential clique family forming a conjugate prior for a family of compatible models.

Let ϑ be a family of compatible distributions for X (such as the family of marginal distributions of a strong hyper Markov law, by Proposition 3.7), with density p with respect to some product measure. Then

$$\pi(X|\mathcal{G}) = \prod_{A \subseteq V} p_A(X_A)^{[t(\mathcal{G})]_A},$$

and thus the posterior law is

$$\pi(\mathcal{G}|X) \propto \exp\left\{\left[\omega + \left(\log p_A(X_A)\right)_{A \subseteq V}\right] \cdot t(\mathcal{G})\right\}.$$

A key benefit of this conjugate formation is that we can describe the posterior law with a parameter of dimension $2^{|V|}$ (strictly speaking, we only need $2^{|V|} - |V| - 1$, due to the over-parametrisation). This is much smaller than for an arbitrary law over the set of undirected decomposable graphs, which would require a parameter of length approximately $2^{\binom{|V|}{2}}$.

4. Ordered directed structural Markov property

We now investigate how the structural Markov property might be extended to directed acyclic graphical models (DAGs). Firstly, we consider a law for a random graph $\tilde{\mathcal{G}}$ over the set \mathfrak{D}^{\prec} : the set of directed acyclic graphs that respect a fixed well ordering \prec on V .

The set \mathfrak{D}^{\prec} is fairly easy to characterise: if an edge exists, its direction is determined by \prec . Furthermore, any subset of the set of pairs of vertices $\binom{V}{2}$ will uniquely characterise a graph in \mathfrak{D}^{\prec} , and therefore

$$|\mathfrak{D}^{\prec}| = 2^{\binom{|V|}{2}}.$$

So how might we develop a structural Markov property for such a graph? Recall that by the strong directed hyper Markov property

$$\tilde{\theta}_{v|\text{pa}(v)} \perp\!\!\!\perp \tilde{\theta}_{\text{pr}(v)}. \quad (4.1)$$

As both $\text{pr}(v)$ and $\text{pr}(v) \cup \{v\}$ are ancestral sets in any such graph, then the projections of the semi-graphoid are equal to those of the induced subgraphs. Furthermore, since the ordered directed Markov property is a spanning set of the semi-graphoid, the semi-graphoid $\mathcal{M}_{\text{pr}(v) \cup \{v\}}(\mathcal{G})$ is spanned by the set

$$\{\langle \{u\}, \text{pr}_{\prec}(u) \mid \text{pa}_{\tilde{\mathcal{G}}}(u) \rangle : u \preceq v\},$$

and hence also by the set

$$\{\langle \{v\}, \text{pr}_{\prec}(v) \mid \text{pa}_{\tilde{\mathcal{G}}}(v) \rangle\} \cup \mathcal{M}(\mathcal{G}_{\text{pr}(v)}). \quad (4.2)$$

Note that $\langle \{v\}, \text{pr}_{\prec}(v) \mid \text{pa}_{\tilde{\mathcal{G}}}(v) \rangle$ only depends on \mathcal{G} through the parent set of v . The correspondence of (4.2) to (4.1) leads to the following definition of an *ordered directed structural Markov property*

$$\text{pa}_{\tilde{\mathcal{G}}}(v) \perp\!\!\!\perp \tilde{\mathcal{G}}_{\text{pr}(v)}.$$

Since this applies for all $v \in V$, we have

$$\coprod_{v \in V} \text{pa}_{\tilde{\mathcal{G}}}(v).$$

As the parent sets of each vertex will uniquely determine the graph, we may easily write the density of such a law as an exponential family whose natural statistic is parent set of each vertex:

$$\pi(\vec{\mathcal{G}}) \propto \exp \left\{ \sum_{v \in V} \sum_{A \subseteq \text{pr}(v)} \omega_{v|A} \mathbb{1}_{\text{pa}_{\vec{\mathcal{G}}}(v)=A} \right\}.$$

5. Markov equivalence and dagoids

Unfortunately the above approach cannot be applied directly to arbitrary directed acyclic graphs. For example, parent sets of individual nodes cannot be independent: if u is a parent of v , then this precludes v from being a parent of u . Before we can define a structural Markov property, we need to explore two key concepts: Markov equivalence and ancestral sets.

Unlike undirected graphs, there is not a one-to-one correspondence between the graph and its semi-graphoid. That is, two or more distinct DAGs may have identical conditional independence properties, as in Figure 2.

Definition 5.1 (Markov equivalence). Let $\vec{\mathcal{G}}$ and $\vec{\mathcal{G}}'$ be directed acyclic graphs such that $\mathcal{M}(\vec{\mathcal{G}}) = \mathcal{M}(\vec{\mathcal{G}}')$. Then $\vec{\mathcal{G}}$ and $\vec{\mathcal{G}}'$ are *Markov equivalent*, which we write as:

$$\vec{\mathcal{G}} \mathcal{M} \vec{\mathcal{G}}'.$$

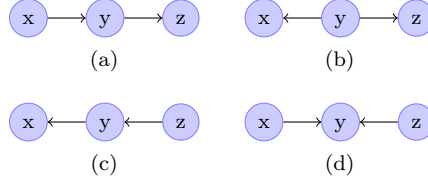


Fig 2: Four directed acyclic graphs with the same skeleton. Graphs (a), (b) and (c) are Markov equivalent, and encode the property $x \perp\!\!\!\perp z \mid y$. Graph (d) has the property $x \perp\!\!\!\perp z$.

So when specifying a law for directed acyclic graphs, we are left with the question of whether or not we should treat Markov equivalent graphs as the same model. In other words, whether the model is defined by the graph or the set of conditional independence statements which it encodes. As noted earlier, we take the latter view.

To simplify notation, we define a *dagoid* to be a Markov equivalence class of directed acyclic graphs. Furthermore, we can define the *complete* and *sparse* dagoids to be the Markov equivalence classes of a complete and sparse DAGs, respectively. We will define $\mathfrak{D}^{\mathcal{M}}$ to be the set of dagoids on V .

A further advantage of working with equivalence classes is that a smaller number of models need be considered. However this may not be as beneficial as one may initially hope: Castelo and Kočka (2004) observed empirically that the ratio of the number DAGs to the number of equivalence classes appears to converge to approximately 3.7, as the number of vertices increases.

5.1. Characterising Markov equivalence

Numerous techniques have been developed for determining whether or not two graphs are Markov equivalent.

5.1.1. Skeleton and immoralities

The *skeleton* of a DAG is the undirected graph obtained by substituting the directed edges for undirected ones. A triplet (a, b, c) of vertices is an *immorality* of a DAG $\vec{\mathcal{G}}$ if the induced graph $\vec{\mathcal{G}}_{\{a,b,c\}}$ is of the form $a \rightarrow b \leftarrow c$.

Theorem 5.1 (Verma and Pearl 1990, Theorem 1; Verma and Pearl 1992, Corollary 3.2; Frydenberg 1990, Theorem 5.6; Andersson, Madigan and Perlman 1997a, Theorem 2.1). *Directed acyclic graphs $\vec{\mathcal{G}}$ and $\vec{\mathcal{G}}'$ are Markov equivalent if and only if they have the same skeleton and the same immoralities.*

5.1.2. Essential graphs

An edge of a DAG $\vec{\mathcal{G}}$ is *essential* if it has the same direction in all Markov equivalent DAGs. The *essential graph* of $\vec{\mathcal{G}}$ is the graph in which all non-essential

edges are replaced by undirected edges.

Although not explored further in this work, the essential graph is a type of *chain graph*, a class of graphs that may have both directed and undirected edges. For further details on chain graphs, in particular their Markov properties and how they relate to undirected and directed acyclic graphs, see [Frydenberg \(1990\)](#) and [Andersson, Madigan and Perlman \(1997b\)](#).

Theorem 5.2 ([Andersson, Madigan and Perlman 1997a](#), Proposition 4.3). *Directed acyclic graphs $\vec{\mathcal{G}}$ and $\vec{\mathcal{G}}'$ are Markov equivalent if and only if they have the same essential graph.*

Unfortunately, there is no simple criterion for determining whether or not an edge of a given DAG is essential, although [Andersson, Madigan and Perlman \(1997a\)](#) developed an iterative algorithm. This limits their usefulness.

5.1.3. Covered edge reversals

A convenient characterisation of Markov equivalence can be given in terms of edge reversals. An edge $a \rightarrow b$ of a DAG $\vec{\mathcal{G}}$ is *covered* if $\text{pa}(b) = \text{pa}(a) \cup \{a\}$.

Theorem 5.3 ([Chickering 1995](#), Theorem 2). *Directed acyclic graphs $\vec{\mathcal{G}}$ and $\vec{\mathcal{G}}'$ are Markov equivalent if and only if there exists a sequence of DAGs:*

$$\vec{\mathcal{G}} = \vec{\mathcal{G}}_0, \vec{\mathcal{G}}_1, \dots, \vec{\mathcal{G}}_{k-1}, \vec{\mathcal{G}}_k = \vec{\mathcal{G}}'$$

such that each $(\vec{\mathcal{G}}_{i-1}, \vec{\mathcal{G}}_i)$ differ only by the reversal of one covered edge.

As we shall see, this result is particularly useful for identifying properties that are preserved under Markov equivalence, as it is only necessary to show that the property is preserved under a covered edge reversal.

5.1.4. Imsets

Imsets for undirected decomposable graphs were briefly mentioned in section 3.4. This formalism can be extended to directed acyclic graphs. The *standard imset* of a directed acyclic graph $\vec{\mathcal{G}}$ is ([Studený, 2005b](#), Page 135)

$$u_{\vec{\mathcal{G}}} = \delta_V - \delta_\emptyset + \sum_{v \in V} \left[\delta_{\text{pa}_{\vec{\mathcal{G}}}(v)} - \delta_{\text{pa}_{\vec{\mathcal{G}}}(v) \cup \{v\}} \right].$$

Theorem 5.4 ([Studený 2005b](#), Corollary 7.1). *Directed acyclic graphs $\vec{\mathcal{G}}$ and $\vec{\mathcal{G}}'$ are Markov equivalent if and only if $u_{\vec{\mathcal{G}}} = u_{\vec{\mathcal{G}}'}$.*

[Studený and Vomlel \(2009\)](#) gives details of the relationship between the imset and the essential graph of a DAG, and how one may be obtained from the other.

5.2. Ancestral sets and remainder dagoids

Ancestral sets appear frequently in the theory of directed acyclic graphical models, as they exhibit an analogous partitioning property to that of decompositions in undirected decomposable graphs. Notably, the global directed Markov property (2.2) can be defined in terms of ancestral sets.

However ancestral sets are not preserved under Markov equivalence, that is, an ancestral set in one graph $\vec{\mathcal{G}}$ need not be ancestral in another Markov equivalent graph $\vec{\mathcal{G}}'$. For example, in Figure 2, $\{x, y\}$ is ancestral in (a) and (b), but not in (c).

As noted earlier, subgraphs induced by ancestral sets preserve the projection of the semi-graphoid. A somewhat trivial consequence is the following.

Proposition 5.5. *Let $\vec{\mathcal{G}} \stackrel{\mathcal{M}}{\sim} \vec{\mathcal{G}}'$, and $A \subseteq V$ be ancestral in both $\vec{\mathcal{G}}$ and $\vec{\mathcal{G}}'$. Then $\vec{\mathcal{G}}_A \stackrel{\mathcal{M}}{\sim} \vec{\mathcal{G}}'_A$.*

This motivates our definition of an ancestral set for a dagoid.

Definition 5.2. A set $A \subseteq V$ is *ancestral* in a dagoid \mathcal{D} if it is ancestral for some graph $\vec{\mathcal{G}} \in \mathcal{D}$. For any such A , define \mathcal{D}_A , the *subdagoid induced by A* , to be the Markov equivalence class of $\vec{\mathcal{G}}_A$.

We further define $\mathfrak{D}(A) \subseteq \mathfrak{D}^{\mathcal{M}}$ to be the set of dagoids in which A is an ancestral set.

This property is not as strong as the collapsibility property in undirected graphs, in that there can exist non-ancestral sets that also preserve the semi-graphoid of the induced subgraph. For example, in Figure 2 (d), the set $\{x, y\}$ is not ancestral, but induced subgraph preserves the (trivial) semi-graphoid.

However ancestral sets are still quite powerful, in that they can be used to decompose the semi-graphoid.

Definition 5.3. Let $\vec{\mathcal{G}}$ be a directed acyclic graph on V , of which A is an ancestral set, and let $\vec{\mathcal{H}}$ be a directed acyclic graph on A . Then the *insertion of $\vec{\mathcal{H}}$ into $\vec{\mathcal{G}}$* , written

$$\vec{\mathcal{H}} \ltimes \vec{\mathcal{G}},$$

is the directed acyclic graph on V with edge set

$$\mathcal{E}(\vec{\mathcal{H}}) \cup [\mathcal{E}(\vec{\mathcal{G}}) \setminus A^2].$$

In other words, the edges between elements of A are determined by $\vec{\mathcal{H}}$, and all other edges are determined by $\vec{\mathcal{G}}$. This operation preserves Markov equivalence.

Lemma 5.6. *Let $\vec{\mathcal{G}}$ and $\vec{\mathcal{G}}'$ be Markov equivalent graphs in which A is an ancestral set, and $\vec{\mathcal{H}}$ and $\vec{\mathcal{H}}'$ be Markov equivalent graphs on A . Then*

$$\vec{\mathcal{H}} \ltimes \vec{\mathcal{G}} \stackrel{\mathcal{M}}{\sim} \vec{\mathcal{H}}' \ltimes \vec{\mathcal{G}}'$$

Proof. Both graphs must have the same skeleton. Let (a, b, c) be an immorality in $\vec{\mathcal{H}} \ltimes \vec{\mathcal{G}}$. Then if $b \in A$, then (a, b, c) must be an immorality of $\vec{\mathcal{H}}$, and hence also an immorality of $\vec{\mathcal{H}}'$, and so also of $\vec{\mathcal{H}}' \ltimes \vec{\mathcal{G}}'$.

Otherwise if $b \notin A$, and at least one of a or c is not in A , then (a, b, c) must be an immorality of $\vec{\mathcal{G}}$, and hence an immorality of $\vec{\mathcal{G}}'$ and $\vec{\mathcal{H}}' \ltimes \vec{\mathcal{G}}'$.

Finally, if $b \notin A$ and $a, c \in A$, then $\{a, c\}$ must not be an edge in the skeleton $\vec{\mathcal{H}}$, nor an edge in the skeleton of $\vec{\mathcal{H}}'$. Hence it must also be an immorality of $\vec{\mathcal{H}}' \ltimes \vec{\mathcal{G}}'$. \square

Consequently for a dagoid \mathcal{D} with ancestral set A , we can define the *ancestral insertion* of a dagoid \mathcal{K} on A into \mathcal{D} as

$$\mathcal{K} \ltimes \mathcal{D} = [\vec{\mathcal{H}} \ltimes \vec{\mathcal{G}}]_{\mathcal{M}},$$

where $\vec{\mathcal{G}} \in \mathcal{D}$ is a directed acyclic graph with an ancestral set A , and $\vec{\mathcal{H}} \in \mathcal{K}$.

We use this approach to decompose the semi-graphoid of a directed acyclic graph.

Definition 5.4. Let A be an ancestral set of a directed acyclic graph $\vec{\mathcal{G}}$. A directed acyclic graph $\vec{\mathcal{G}}_{V|A}$ is a *remainder graph* of $\vec{\mathcal{G}}$ given A if

$$\vec{\mathcal{G}}_{V|A} = \mathcal{C}^{(A)} \ltimes \vec{\mathcal{G}}$$

where $\mathcal{C}^{(A)}$ is a complete dagoid on A .

By Lemma 5.6, the remainder graph must be unique up to Markov equivalence. Hence for a dagoid $\mathcal{D} \in \mathfrak{D}(A)$, we can uniquely define the *remainder dagoid* of \mathcal{D} given A , denoted by $\mathcal{D}_{V|A}$.

The name comes from the fact that $\mathcal{M}(\mathcal{D}_A)$ and $\mathcal{M}(\mathcal{D}_{V|A})$ form a spanning subset of $\mathcal{M}(\mathcal{D})$.

Theorem 5.7. Let A be an ancestral set of a directed acyclic graph $\vec{\mathcal{G}}$. Then

$$\mathcal{M}(\vec{\mathcal{G}}) = \overline{\mathcal{M}(\vec{\mathcal{G}}_A) \cup \mathcal{M}(\vec{\mathcal{G}}_{V|A})},$$

where \overline{S} denotes the Markov closure of a set of conditional independence statements S .

Proof. Recall that $\mathcal{M}(\vec{\mathcal{G}})$ is spanned by the set of elements of the form:

$$\langle \{v\}, \text{pr}_{\prec}(v) \mid \text{pa}_{\vec{\mathcal{G}}}(v) \rangle \quad (5.1)$$

where \prec is a well-ordering in which the elements of A precede those of $V \setminus A$. If $v \in A$, then (5.1) will be an element of $\mathcal{M}(\vec{\mathcal{G}}_A)$, otherwise if $v \notin A$, it will be an element of $\mathcal{M}(\vec{\mathcal{G}}_{V|A})$. \square

Furthermore, the induced and remainder dagoids are variation independent.

Theorem 5.8. For any $A \subseteq V$, we have:

$$\mathcal{D}_A \nmid \mathcal{D}_{V|A} \mid \{\mathcal{D} \in \mathfrak{D}(A)\}$$

Proof. For any $\mathcal{D}, \mathcal{D}' \in \mathfrak{D}A$, we can construct $\mathcal{D}^* = \mathcal{D}_A \ltimes \mathcal{D}'_{V|A}$. This will have the required properties that $\mathcal{D}_A^* = \mathcal{D}_A$ and $\mathcal{D}_{V|A}^* = \mathcal{D}'_{V|A}$. \square

6. Dagoid structural Markov property

Recall that the strong hyper Markov property for the law $\mathcal{L}(\tilde{\theta})$ can be expressed as

$$\bigsqcup_{v \in V} \tilde{\theta}_{v|\text{pa}(v)} \quad [\mathcal{L}].$$

For any ancestral set A of $\vec{\mathcal{G}}$, we can write

$$\theta_A \simeq (\theta_{v|\text{pa}(v)})_{v \in A} \quad \text{and} \quad \theta_{V|A} \simeq (\theta_{v|\text{pa}(v)})_{v \notin A}.$$

Therefore, an alternative characterisation of the strong hyper Markov property is

$$\tilde{\theta}_A \perp\!\!\!\perp \tilde{\theta}_{V|A} \quad [\mathcal{L}],$$

for any ancestral set A of $\vec{\mathcal{G}}$.

This motivates the following definition:

Definition 6.1 (Dagoid structural Markov property). We say a graph law $\mathfrak{G}(\tilde{\mathcal{D}})$ is *structurally Markov* if for any $A \subseteq V$, we have

$$\tilde{\mathcal{D}}_{V|A} \perp\!\!\!\perp \tilde{\mathcal{D}}_A \mid \{\tilde{\mathcal{D}} \in \mathfrak{D}(A)\} \quad [\mathfrak{G}].$$

As in the undirected case, we can characterise this property via the odds ratio of the density:

Proposition 6.1. *A graph law is structurally Markov if and only if for any $\mathcal{D}, \mathcal{D}' \in \mathfrak{D}(A)$, we have:*

$$\pi(\mathcal{D})\pi(\mathcal{D}') = \pi(\mathcal{D}_A \ltimes \mathcal{D}'_{V|A})\pi(\mathcal{D}'_A \ltimes \mathcal{D}_{V|A}). \quad (6.1)$$

Proof. As in Proposition 3.2, we may write the density $\pi(\mathcal{D} \mid \mathfrak{D}(A)) = \pi(\mathcal{D}_A \mid \mathfrak{D}(A))\pi(\mathcal{D}_{V|A} \mid \mathfrak{D}(A))$. \square

Example 6.1. As in the undirected case, the simplest example of a structurally Markov graph law is the uniform law over $\mathfrak{D}^{\mathcal{M}}$.

However, we note that some simple laws are *not* structurally Markov.

Example 6.2. Consider the law in which $\pi(\mathcal{D})$ is proportional to $|\mathcal{D}|$, in other words, the uniform law on \mathfrak{D} projected onto $\mathfrak{D}^{\mathcal{M}}$. Then we note the size of the following dagoids:

$$\begin{aligned} [\text{graph}]_{\mathcal{M}} &= \{ \text{graph} \} \\ [\text{graph}]_{\mathcal{M}} &= \{ \text{graph}, \text{graph} \} \\ [\text{graph}]_{\mathcal{M}} &= \{ \text{graph} \} \\ [\text{graph}]_{\mathcal{M}} &= \{ \text{graph}, \text{graph}, \text{graph}, \text{graph}, \text{graph}, \text{graph} \} \end{aligned}$$

As a consequence, this law doesn't satisfy Proposition 6.1, when $\mathcal{D} = \text{graph}$ and $\mathcal{D}' = \text{graph}$, and A is chosen as the two top vertices.

6.1. *d*-Clique vector

The equivalence class formulation of a dagoid is difficult to work with, both algebraically and computationally. Instead we propose a characteristic vector similar to the clique vector of Section 3.4.

Definition 6.2. The *d-clique vector* of a directed acyclic graph $\vec{\mathcal{G}}$ is:

$$t(\vec{\mathcal{G}}) = \sum_{v \in V} [\delta_{\{v\} \cup \text{pa}_{\vec{\mathcal{G}}}(v)} - \delta_{\text{pa}_{\vec{\mathcal{G}}}(v)}] + \delta_{\emptyset} \in \mathbb{Z}^{2^V}, \quad (6.2)$$

where

$$(\delta_A)_I = \begin{cases} 1 & \text{if } I = A \\ 0 & \text{if } I \neq A \end{cases}.$$

Again, we note the relationship to the imsets of Studený (2005b), specifically the structural imset $t(\vec{\mathcal{G}}) = \delta_V - u_{\vec{\mathcal{G}}}$ of in section 5.1.4

In a similar manner to the undirected case, we can define the *d-completeness vector* to be the Möbius transform of the d-clique vector,

$$c_A(\vec{\mathcal{G}}) = \sum_{B \supseteq A} t_B(\vec{\mathcal{G}}), \quad (6.3)$$

and say that a set $A \subseteq B$ is *d-complete* if $c_A(\vec{\mathcal{G}}) = 1$.

Lemma 6.2. Let \prec be a well-ordering of a directed acyclic graph $\vec{\mathcal{G}}$. Then for any non-empty set $A \subseteq V$:

$$c_A(\vec{\mathcal{G}}) = \begin{cases} 1 & \text{if } A \setminus \{a\} \subseteq \text{pa}_{\vec{\mathcal{G}}}(a), \\ 0 & \text{otherwise,} \end{cases}$$

where a is the maximal element of A under \prec .

Proof. By substituting (6.2) into (6.3):

$$c_A(\vec{\mathcal{G}}) = \sum_{v \in V} \mathbb{1}_{A \subseteq \text{pa}_{\vec{\mathcal{G}}}(v) \cup \{v\}} - \mathbb{1}_{A \subseteq \text{pa}_{\vec{\mathcal{G}}}(v)}.$$

These terms will cancel out unless $v \in A$. Furthermore, $A \subseteq \text{pa}_{\vec{\mathcal{G}}}(v) \cup \{v\}$ only if all $u \prec v$ for all $u \in A$. Hence:

$$c_A(\vec{\mathcal{G}}) = \mathbb{1}_{A \subseteq \text{pa}_{\vec{\mathcal{G}}}(a) \cup \{a\}}. \quad \square$$

This provides the link to the completeness and clique vectors of undirected graphs from section 3.4.

Corollary 6.3. If $\vec{\mathcal{G}}$ is a perfect directed acyclic graph and $\vec{\mathcal{G}}^s$ is its skeleton, then $c_{\vec{\mathcal{G}}} = c_{\vec{\mathcal{G}}^s}$, and hence $t(\vec{\mathcal{G}}) = t(\vec{\mathcal{G}}^s)$.

Most importantly, the d-clique vector is a unique representation of the dagoid.

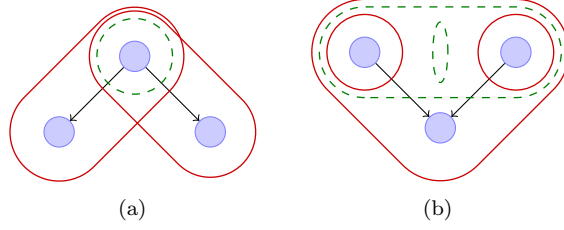


Fig 3: The d-cliques (—) and d-separators (---) of different directed acyclic graphs. Note that in the perfect DAG (a), the d-cliques and d-separators are the cliques and separators of the skeleton. However, as in (b), d-separators may contain d-cliques.

Theorem 6.4. *Let $\vec{\mathcal{G}}, \vec{\mathcal{G}}'$ be directed acyclic graphs on V . Then $\vec{\mathcal{G}} \stackrel{\mathcal{M}}{\sim} \vec{\mathcal{G}}'$ if and only if $t(\vec{\mathcal{G}}) = t(\vec{\mathcal{G}}')$.*

Proof. To show the d-clique vector is preserved under Markov equivalence, by Theorem 5.3 it is sufficient to show that it is preserved under a covered edge reversal. If (a, b) is a covered edge of $\vec{\mathcal{G}}$, then the contribution of these vertices to the sum (6.2) is:

$$\begin{aligned} t(\vec{\mathcal{G}}) &= [\delta_{\{a\} \cup \text{pa}_{\vec{\mathcal{G}}}(a)} - \delta_{\text{pa}_{\vec{\mathcal{G}}}(a)}] + [\delta_{\{b\} \cup \text{pa}_{\vec{\mathcal{G}}}(b)} - \delta_{\text{pa}_{\vec{\mathcal{G}}}(b)}] \\ &\quad + \sum_{v \neq a, b} [\delta_{\{b\} \cup \text{pa}_{\vec{\mathcal{G}}}(b)} - \delta_{\text{pa}_{\vec{\mathcal{G}}}(b)}] + \delta_{\emptyset} \end{aligned}$$

By definition $\text{pa}_{\vec{\mathcal{G}}}(a) \cup \{a\} = \text{pa}_{\vec{\mathcal{G}}}(b)$, and so the corresponding terms will cancel. If $\vec{\mathcal{G}}^*$ is obtained from $\vec{\mathcal{G}}$ by reversing (a, b) , note that:

$$\text{pa}_{\vec{\mathcal{G}}}(a) = \text{pa}_{\vec{\mathcal{G}}^*}(b) \quad \text{and} \quad \text{pa}_{\vec{\mathcal{G}}}(b) \cup \{b\} = \text{pa}_{\vec{\mathcal{G}}^*}(a) \cup \{a\},$$

and the remaining terms will be unchanged. Hence $t(\vec{\mathcal{G}}) = t(\vec{\mathcal{G}}^*)$.

To show that the d-completeness vector (and hence, also the d-clique vector) is unique to the equivalence class, by Theorem 5.1 we can show that it determines the skeleton and immoralities. By Lemma 6.2, there is an edge between u and v in $\vec{\mathcal{G}}$ if and only if $c_{\{u, v\}}(\vec{\mathcal{G}}) = 1$. Likewise, (u, v, w) is an immorality if and only if $c_{\{u, v, w\}}(\vec{\mathcal{G}}) = 1$ and $c_{\{u, w\}}(\vec{\mathcal{G}}) = 0$. \square

This cancellation of terms involving covered edges is very useful: as a consequence, the d-clique vector will generally be quite sparse. In line with the clique vector, we term sets $A \subseteq V$ such that $t_A(\mathcal{D}) = 1$ to be a *d-clique*, and the sets where $t_A(\mathcal{D}) < 0$ to be *d-separators*. See examples in Figure 3.

Theorem 6.5. *Let A be an ancestral set of a dagoid \mathcal{D} . Then*

$$t(\mathcal{D}) = [t(\mathcal{D}_A)]^0 + t(\mathcal{D}_{V|A}) - \delta_A,$$

where $[\cdot]^0$ denotes the expansion of the vector with zeroes to the required coordinates.

Proof. Let $\vec{\mathcal{G}} \in \mathcal{D}$ in which A is ancestral, and \prec be a well-ordering of $\vec{\mathcal{G}}$ in which elements of A precede those of $V \setminus A$. Then

$$\text{pa}_{\vec{\mathcal{G}}}(v) = \begin{cases} \text{pa}_{\vec{\mathcal{G}}_A}(v) & v \in A, \\ \text{pa}_{\vec{\mathcal{G}}_{V|A}}(v) & v \notin A. \end{cases}$$

The result follows after noting that

$$\sum_{v \in A} [\delta_{(\text{pa}_{\vec{\mathcal{G}}_{V|A}}(v) \cup \{v\})} - \delta_{\text{pa}_{\vec{\mathcal{G}}_{V|A}}(v)}] = \delta_A. \quad \square$$

We now arrive at the key result of this section: the dagoid structural Markov property characterises an exponential family of graph laws.

Theorem 6.6. *Let \mathfrak{G} whose support is $\mathfrak{D}^{\mathcal{M}}$. Then \mathfrak{G} is structurally Markov if and only if it is a member of an exponential family with the d -clique vector as a sufficient statistic. That is, \mathfrak{G} has density*

$$\pi_{\omega}(\mathcal{D}) \propto \exp\{\omega \cdot t(\mathcal{D})\}. \quad (6.4)$$

Proof. If the law is in the exponential family in (6.4), then by Theorem 6.5

$$\pi(\mathcal{D}|\mathfrak{D}(A)) \propto \exp\{\omega \cdot [t(\mathcal{D}_A) + t(\mathcal{D}_{V|A})] - \omega_A\} \propto p(\mathcal{D}_A|\mathfrak{D}(A))p(\mathcal{D}_{V|A}|\mathfrak{D}(A)),$$

and hence the law must be structurally Markov.

For the converse, define $\mathcal{D}^{(A)}$ to be the dagoid in which the induced dagoid on $A \subseteq V$ is complete, but otherwise sparse (in other words, the remainder dagoid $\mathcal{D}_{V|A}^{(\emptyset)}$, of the sparse dagoid $\mathcal{D}^{(\emptyset)}$).

Select some $\vec{\mathcal{G}} \in \mathcal{D}$, and let v_1, \dots, v_d be a well ordering of V . Recursively define the dagoids:

$$\mathcal{D}^{*(i)} = \begin{cases} \mathcal{D}(\{v_1\}) & \text{if } i = 1, \\ \mathcal{D}_{\text{pr}(v_i)}^{*(i-1)} \ltimes \mathcal{D}_{v_i|\text{pr}(v_i)}^{(\{v_i\} \cup \text{pa}(v_i))} & \text{otherwise.} \end{cases}$$

By Proposition 6.1, for $i = 2, \dots, d$

$$\pi(\mathcal{D}^{*(i-1)})\pi(\mathcal{D}^{(\{v_i\} \cup \text{pa}(v_i))}) = \pi(\mathcal{D}^{*(i)})\pi(\mathcal{D}_{\text{pr}(v_i)}^{(\{v_i\} \cup \text{pa}(v_i))} \ltimes \mathcal{D}_{v_i|\text{pr}(v_i)}^{*(i-1)})$$

However,

$$\mathcal{D}_{\text{pr}(v_i)}^{(\{v_i\} \cup \text{pa}(v_i))} \ltimes \mathcal{D}_{v_i|\text{pr}(v_i)}^{*(i-1)} = \mathcal{D}^{(\text{pa}(v_i))}$$

Therefore, since $\mathcal{D}^{*(d)} = \mathcal{D}$, then

$$\pi(\mathcal{D}) = \frac{\prod_{i=1}^d \pi(\mathcal{D}^{(\{v_i\} \cup \text{pa}(v_i))})}{\prod_{i=2}^d \pi(\mathcal{D}^{(\text{pa}(v_i))})}.$$

which is of the form in (6.4), where

$$\omega_A = \log \pi(\mathcal{D}^{(A)}). \quad \square$$

We note that a similar exponential families were proposed by [Mukherjee and Speed \(2008\)](#), however they treat Markov equivalent graphs as distinct, and allow them to have different probabilities.

6.2. Compatible distributions and laws

As with the undirected case, a graph law is only part of the story. For each dagoid \mathcal{D} , we also require a method to specify either a Markov sampling distribution, or a law over such sampling distributions.

Definition 6.3. Distributions θ and θ' which are Markov with respect to directed acyclic graphs $\vec{\mathcal{G}}$ and $\vec{\mathcal{G}}'$, respectively, are *graph compatible* if for every vertex v where $\text{pa}_{\vec{\mathcal{G}}}(v) = \text{pa}_{\vec{\mathcal{G}}'}(v)$, there exists versions of the conditional probability distributions for $X_v \mid X_{\text{pa}(v)}$ such that:

$$\theta(X_v \mid X_{\text{pa}(v)}) = \theta'(X_v \mid X_{\text{pa}(v)}).$$

Furthermore, distributions θ and θ' which are Markov with respect to dagoids \mathcal{D} and \mathcal{D}' , respectively, are (*dagoid*) *compatible* if they are graph compatible for every pair of graphs $\vec{\mathcal{G}} \in \mathcal{D}, \vec{\mathcal{G}}' \in \mathcal{D}'$.

Likewise, laws $\mathcal{L}(\tilde{\theta})$ and $\mathcal{L}'(\tilde{\theta})$, hyper Markov with respect to $\vec{\mathcal{G}}$ and $\vec{\mathcal{G}}'$ respectively, are *graph hyper compatible* if for every vertex v where $\text{pa}_{\vec{\mathcal{G}}}(v) = \text{pa}_{\vec{\mathcal{G}}'}(v)$, there exists versions of the conditional laws for $\tilde{\theta}_{v \mid \text{pa}(v)} \mid \tilde{\theta}_{\text{pa}(v)}$ such that:

$$\mathcal{L}(\tilde{\theta}_{v \mid \text{pa}(v)} \mid \tilde{\theta}_{\text{pa}(v)}) = \mathcal{L}'(\tilde{\theta}_{v \mid \text{pa}(v)} \mid \tilde{\theta}_{\text{pa}(v)}).$$

By [Dawid \(2001, section 8.2\)](#), the weak hyper Markov property may be characterised in terms of $\mathcal{M}(\vec{\mathcal{G}})$, and so the weak hyper Markov property can be defined with respect to a dagoid. Laws $\mathcal{L}(\tilde{\theta})$ and $\mathcal{L}'(\tilde{\theta})$, that are hyper Markov with respect to \mathcal{D} and \mathcal{D}' , respectively, are (*dagoid*) *hyper compatible* if they are graph compatible for every pair of graphs $\vec{\mathcal{G}} \in \mathcal{D}, \vec{\mathcal{G}}' \in \mathcal{D}'$.

As in the undirected case, we can define a family of compatible distributions $\vartheta = \{\theta^{(\mathcal{G})} : \mathcal{G} \in \mathfrak{U}\}$ and a family of hyper compatible laws $\mathfrak{L} = \{\mathcal{L}^{(\mathcal{G})} : \mathcal{G} \in \mathfrak{U}\}$ if they are pairwise compatible or hyper compatible with respect to the relevant graphs.

Proposition 6.7. Suppose $\mathfrak{G}(\tilde{\mathcal{D}})$ is a graph law over $\mathfrak{D}^{\mathcal{M}}$ and ϑ is a family of compatible distributions. Then:

$$X_A \perp\!\!\!\perp \tilde{\mathcal{D}}_{V \setminus A} \mid \tilde{\mathcal{D}}_A, \{\tilde{\mathcal{D}} \in \mathfrak{D}(A)\} \quad [\vartheta, \mathfrak{G}] \quad (6.5)$$

and

$$X_{V \setminus A} \perp\!\!\!\perp \tilde{\mathcal{D}}_A \mid X_A, \tilde{\mathcal{D}}_{V \setminus A}, \{\tilde{\mathcal{D}} \in \mathfrak{D}(A)\} \quad [\vartheta, \mathfrak{G}]. \quad (6.6)$$

Likewise, if $\mathfrak{G}(\tilde{\mathcal{D}})$ is a graph law over $\mathfrak{D}^{\mathcal{M}}$ and \mathfrak{L} is a hyper compatible family of laws, then:

$$\tilde{\theta}_A \perp\!\!\!\perp \tilde{\mathcal{D}}_{V \setminus A} \mid \tilde{\mathcal{D}}_A, \{\tilde{\mathcal{D}} \in \mathfrak{D}(A)\} \quad [\mathfrak{L}, \mathfrak{G}]$$

and

$$\tilde{\theta}_{V \setminus A|A} \perp\!\!\!\perp \tilde{\mathcal{D}}_A \mid \tilde{\theta}_A, \tilde{\mathcal{D}}_{V|A}, \{\tilde{\mathcal{D}} \in \mathfrak{D}(A)\} \quad [\mathfrak{L}, \mathfrak{G}].$$

Proof. This is much the same as Proposition 3.8: for (6.5), the distribution for X_A are determined by the parent sets of the vertices in A in some $\mathcal{G} \in \mathcal{D}$ in which A is ancestral. Likewise, in (6.6), the conditional distribution for $X_{V \setminus A} \mid X_A$ is determined by the parents sets of vertices in $V \setminus A$. The same argument applies at the hyper level. \square

Note that in the definition of compatibility and hyper compatibility we specifically refer to *versions* of conditional probabilities and laws, as in some cases the conditional distributions/laws will not be uniquely defined, due to conditioning on null sets.

Example 6.3. Suppose the joint distribution P on a triplet of binary variables (X, Y, Z) has $P(X = 1, Y = 1) = 0$, but with $P(X = 1) > 0$ and $P(Y = 1) > 0$. Then the conditional distribution $P(Z = 1 \mid X = 1, Y = 1)$ is not uniquely defined.

Now consider a compatible distribution P' on the graph:



Then we have $P'(X = 1, Y = 1) = P(X = 1)P(Y = 1) > 0$. Therefore $P'(X = 1, Y = 1, Z = 1)$ may be defined arbitrarily, as for any conditional probability $P'(Z = 1 \mid X = 1, Y = 1)$, there will exist a corresponding version of $P(Z = 1 \mid X = 1, Y = 1)$.

We could avoid this type of ambiguity in the case of compatible distributions by requiring that the density be positive with respect to some product measure. However the situation isn't so simple at the hyper level:

Example 6.4. Consider a law $\mathcal{L}(\tilde{\theta})$ for a triplet of binary variables (X, Y, Z) , and suppose that it is continuous on the full probability simplex.

A hyper compatible law \mathcal{L}' on the graph in (6.7), will have marginal laws $\mathcal{L}'(\tilde{\theta}_X) = \mathcal{L}(\tilde{\theta}_X)$ and $\mathcal{L}'(\tilde{\theta}_Y) = \mathcal{L}(\tilde{\theta}_Y)$. This means the joint law $\mathcal{L}'(\tilde{\theta}_{XY})$ will be their product law, which is concentrated on the manifold $X \perp\!\!\!\perp Y$.

As this manifold will have probability 0 under \mathcal{L} , we may define the conditional law $\mathcal{L}'(\tilde{\theta}_{Z|XY} \mid \tilde{\theta}_{XY})$ arbitrarily.

It is possible to uniquely define such conditional laws if we impose further conditions, such as the existence of a continuous density for $\mathcal{L}(\tilde{\theta})$. However we can also resolve the problem by insisting on a dagoid form of the strong hyper Markov property:

Definition 6.4. A law $\mathcal{L}(\tilde{\theta})$ is over $\mathfrak{P}(\mathcal{D})$ is *strong hyper Markov* with respect to \mathcal{D} if it is strong directed hyper Markov with respect to every $\vec{\mathcal{G}} \in \mathcal{D}$.

Note that if $\vec{\mathcal{G}} \in \mathcal{D}$ is perfect, then the dagoid strong hyper Markov property is equivalent to the undirected strong hyper Markov property on the skeleton of $\vec{\mathcal{G}}$ (see (Dawid and Lauritzen, 1993, Proposition 3.15)).

The notion of hyper compatibility is equivalent to the “parameter modularity” property of [Heckerman, Geiger and Chickering \(1995\)](#). Likewise, the strong hyper Markov property is equivalent to their “parameter independence”

Example 6.5 (Dagoid hyper inverse Wishart law). For each vertex v of a directed acyclic graph $\vec{\mathcal{G}}$, we define the law for the conditional parameter $\mathcal{L}(\tilde{\theta}_v | \text{pa}_{\vec{\mathcal{G}}}(v))$ to be the same as that of the inverse Wishart $\mathcal{IW}(\delta; \Phi)$. That is, using the notation of [Dawid \(1981\)](#),

$$\begin{aligned} \mathcal{L}(\tilde{\Sigma}_v | \text{pa}_{\vec{\mathcal{G}}}(v)) &= \mathcal{IW}(\delta + |\text{pa}_{\vec{\mathcal{G}}}(v)|; \Phi_{v | \text{pa}_{\vec{\mathcal{G}}}(v)}) \\ \mathcal{L}(\tilde{\Gamma}_{v | \text{pa}_{\vec{\mathcal{G}}}(v)} | \tilde{\Sigma}_v | \text{pa}_{\vec{\mathcal{G}}}(v)) &= \Phi_{\{v\}, \text{pa}_{\vec{\mathcal{G}}}(v)} \Phi_{\text{pa}_{\vec{\mathcal{G}}}(v)}^{-1} + \mathcal{N}_{\{v\} \times \text{pa}_{\vec{\mathcal{G}}}(v)}(\tilde{\Sigma}_v | \text{pa}_{\vec{\mathcal{G}}}(v), \Phi_{\text{pa}_{\vec{\mathcal{G}}}(v)}^{-1}) \end{aligned}$$

By the properties of the inverse Wishart law, it follows that the law derived under a covered edge reversal will be identical, hence may be defined by the dagoid. Furthermore, by the above definition, it is hyper compatible.

Theorem 6.8. *If \mathfrak{L} is a family of strong hyper Markov hyper compatible laws, then the family of marginal data distributions is compatible*

Proof. The hyper compatibility and the strong hyper Markov property imply that for any two dagoids $\mathcal{D}, \mathcal{D}'$, and any $\vec{\mathcal{G}} \in \mathcal{D}, \vec{\mathcal{G}}' \in \mathcal{D}'$, that if $\text{pa}_{\vec{\mathcal{G}}}(v) = \text{pa}_{\vec{\mathcal{G}}'}(v)$ for some $v \in V$, then:

$$\mathcal{L}^{(\mathcal{D})}(\tilde{\theta}_v | \text{pa}) = \mathcal{L}^{(\mathcal{D}')}(\tilde{\theta}_v | \text{pa})$$

Therefore, the family of marginal data distributions $\bar{\vartheta} = \{\bar{\theta}^{(\mathcal{D})} : \mathcal{D} \in \mathfrak{D}^{\mathcal{M}}\}$ will have:

$$\bar{\theta}^{(\mathcal{D})}(X_v | X_{\text{pa}_{\vec{\mathcal{G}}}}) = \mathbb{E}_{\mathcal{L}^{(\mathcal{D})}}[\tilde{\theta}_v | \text{pa}_{\vec{\mathcal{G}}}] = \bar{\theta}^{(\mathcal{D}')} (X_v | X_{\text{pa}_{\vec{\mathcal{G}}'}}) = \mathbb{E}_{\mathcal{L}^{(\mathcal{D}')}}[\tilde{\theta}_v | \text{pa}_{\vec{\mathcal{G}}}]. \quad \square$$

This is particularly useful because, as in the undirected case, the structural Markov property will be preserved in the posterior under compatible sampling:

Theorem 6.9. *Suppose $\mathfrak{G}(\tilde{\mathcal{D}})$ is a structurally Markov graph law over $\mathfrak{D}^{\mathcal{M}}$ and ϑ is a family of compatible distributions. Then the posterior graph law for $\tilde{\mathcal{D}}$ is structurally Markov.*

Proof. By the structural Markov property and (6.5), we have:

$$(X_A, \tilde{\mathcal{D}}_A) \perp\!\!\!\perp \tilde{\mathcal{D}}_{V|A} \mid \{\tilde{\mathcal{D}} \in \mathfrak{D}(A)\}$$

and hence:

$$\tilde{\mathcal{D}}_A \perp\!\!\!\perp \tilde{\mathcal{D}}_{V|A} \mid X_A, \{\tilde{\mathcal{D}} \in \mathfrak{D}(A)\}$$

Combining this with (6.6), we get:

$$\tilde{\mathcal{D}}_A \perp\!\!\!\perp (\tilde{\mathcal{D}}_{V|A}, X_{V \setminus A}) \mid X_A, \{\tilde{\mathcal{D}} \in \mathfrak{D}(A)\}$$

and hence:

$$\tilde{\mathcal{D}}_A \perp\!\!\!\perp \tilde{\mathcal{D}}_{V|A} \mid X, \{\tilde{\mathcal{D}} \in \mathfrak{D}(A)\} \quad \square$$

We can specify a compatible family by a positive density on the complete dagoid:

Theorem 6.10. *If the distribution on the complete dagoid has positive density p (with respect to some product measure), then the compatible distribution for any dagoid \mathcal{D} , has density:*

$$p^{(\mathcal{D})}(x) = \prod_{A \subseteq V} [p(x_A)]^{t(\mathcal{D})_A} \quad (6.8)$$

Proof. Let $\vec{\mathcal{G}}$ be an arbitrary graph in \mathcal{D} . Then by compatibility:

$$p^{(\mathcal{D})}(x) = \prod_{v \in V} p(x_v | x_{\text{pa}(v)}) = \frac{\prod_{i=1}^p p(x_{\{v_i\} \cup \text{pa}(v_i)})}{\prod_{i=2}^p p(x_{\text{pa}(v_i)})} = \prod_{A \subseteq V} [p(x_A)]^{t(\mathcal{D})_A} \quad \square$$

As a consequence, if the graph law has a d -clique exponential family of the form (6.4), and the sampling distributions are compatible with density of the form (6.8), then the posterior graph law will have density:

$$\pi(\mathcal{D} | X) \propto \exp\{[\omega + (\log p_A(X_A))_{A \subseteq V}] \cdot t(\mathcal{D})\}.$$

That is, the d -clique exponential family is a conjugate prior under sampling from a compatible family.

7. Computing with structural Markov laws

For even small numbers of vertices it can quickly become infeasible to enumerate all graphs, and hence some sort of numerical approximation will usually be required. Markov chain Monte Carlo (MCMC) is commonly utilised in statistics for precisely this purpose. We provide an overview of various MCMC techniques for these graphical structures, and how they might be adapted for use with structural Markov laws.

7.1. Undirected decomposable graphs

A common approach to constructing MCMC algorithms on graphs relies on small perturbations to the edge set of the graph. The simplest algorithm, proposed by Giudici and Green (1999), relies on making single edge additions and removals. However a key difficulty with such an approach is to characterise which edge modifications will result in the graph remaining decomposable.

For any graph $\mathcal{G} \in \mathcal{U}$, we define $\mathcal{N}^-(\mathcal{G})$ and $\mathcal{N}^+(\mathcal{G})$ to be the set of undirected decomposable graphs that may be obtained by removing or adding, respectively, a single edge from \mathcal{G} . We call these the *lower* and *upper neighbours* of \mathcal{G} .

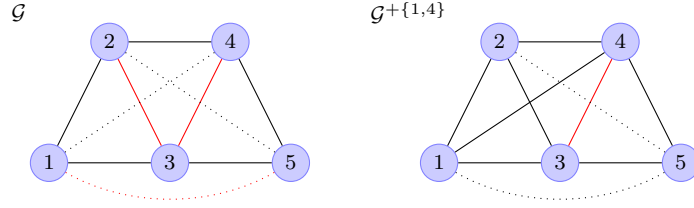


Fig 4: Neighbouring graphs on 5 vertices: solid lines (—) indicate edges, dotted lines (.....) for missing edges. Red lines (—.....) are those whose removal/addition will result in a non-decomposable graph. In \mathcal{G} , only 7 of the 10 edges may be modified, whereas in $\mathcal{G}^{+\{1,4\}}$, obtained by adding the edge $\{1,4\}$, 9 of the 10 edges may be modified.

Frydenberg and Lauritzen (1989, Lemma 3) showed that the graph $\mathcal{G}^{-\{u,v\}}$ obtained by removing $\{u,v\}$ is decomposable if and only if $\{u,v\}$ is a subset of exactly one clique C of \mathcal{G} . As a consequence, the set of lower neighbours $\mathcal{N}^{-}(\mathcal{G})$ can be partitioned according to the clique of \mathcal{G} which contained the removed edge. Moreover, for such an edge removal the change in clique vector t is

$$t_A(\mathcal{G}^{-\{u,v\}}) - t_A(\mathcal{G}) = \begin{cases} -1 & \text{if } A = C \text{ or } C \setminus \{u,v\}, \\ +1 & \text{if } A = C \setminus \{u\} \text{ or } C \setminus \{v\}, \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, (Giudici and Green, 1999, Theorem 2) characterises the set of possible edge additions: the graph $\mathcal{G}^{+\{u,v\}}$ obtained by the addition of the edge $\{u,v\}$ is decomposable if and only if there exist cliques (of \mathcal{G}) $C_u \ni u$ and $C_v \ni v$ such that $S = C_u \cap C_v$ is a separator of C_u and C_v in \mathcal{G} . Consequently, the set of upper neighbours $\mathcal{N}^{+}(\mathcal{G})$ can be partitioned according to the separators of \mathcal{G} by which they are separated. Consequently, the change in t is simply

$$t_A(\mathcal{G}^{+\{u,v\}}) - t_A(\mathcal{G}) = \begin{cases} +1 & \text{if } A = S \text{ or } S \cup \{u,v\}, \\ -1 & \text{if } A = S \cup \{u\} \text{ or } S \cup \{v\}, \\ 0 & \text{otherwise.} \end{cases}$$

These sets of neighbouring graphs are illustrated in figure 4.

These criteria can be used to construct an MCMC algorithm for sampling from a structurally Markov graph law, based on that of Giudici and Green (1999). Specifically, we can construct a Metropolis–Hastings algorithm with the following transition kernel: given our current graph $\mathcal{G}^{(t)}$, we select a pair of distinct vertices $u, v \in V$:

- If $\{u, v\} \in \mathcal{E}(\mathcal{G}^{(t)})$, and $\mathcal{G}^{-\{u,v\}}$ is decomposable, then set $\mathcal{G}^{(t+1)} = \mathcal{G}^{-\{u,v\}}$ with probability

$$\min(\exp\{\omega_{C \setminus \{u\}} + \omega_{C \setminus \{v\}} - \omega_{C \setminus \{u,v\}} - \omega_C\}, 1).$$

- If $\{u, v\} \notin \mathcal{E}(\mathcal{G}^{(t)})$, and $\mathcal{G}^{+\{u,v\}}$ is decomposable, then set $\mathcal{G}^{(t+1)} = \mathcal{G}^{+\{u,v\}}$ with probability

$$\min(\exp\{\omega_S + \omega_{S \cup \{u,v\}} - \omega_{S \cup \{u\}} - \omega_{S \cup \{v\}}\}, 1).$$

- Otherwise, set $\mathcal{G}^{(t+1)} = \mathcal{G}^{(t)}$.

This means that at each step, the acceptance probability can be evaluated locally, utilising only four elements of the parameter vector: this is particularly useful when sampling from a posterior distribution (see section 3.6), as we only then need to evaluate the marginal likelihood of four subsets of V .

As it is possible to move between any two decomposable graphs by a sequence of edge additions and removals (Frydenberg and Lauritzen, 1989, Lemma 5), the algorithm is ergodic, and will have the desired invariant density.

One practical issue is the construction of an appropriate data structure to represent the graph in computer memory. It is far from obvious how to efficiently determine if a proposed edge satisfies these criteria. It is worth noting that simply storing a graph as a set of vertices and edges is clearly inefficient, as this would require recomputing the cliques at each step. The results of Thomas and Green (2009a,b) indicate that a list of cliques stored in a perfect sequence or some representation of a clique tree could be useful for this purpose.

Another problem is the rate of mixing of the Markov chain. Due to the extremely large size of the space \mathcal{U} and the restriction on staying within the space of decomposable graphs, it can take an extremely long time to transition between two graphs. Kijima et al. (2007, 2008) show that for a uniform graph law, certain starting graphs will result in a mixing time exponential in $|V|$.

One possible solution is to propose larger jumps. Green and Thomas (2013) suggest an extension of the above scheme in which multiple edges may be removed or added, resulting in a chain that is able to make bolder moves. This algorithm is also able to take advantage of local computations in computing the acceptance ratio.

Another alternative would be to completely separate a vertex from the graph and reconnect it in some other way. However as the sample space for such a proposal scheme would be considerably larger— $|V| \times 2^{|V|-1}$ instead of $\binom{|V|}{2}$ —a uniform proposal distribution could result in frequently proposing moves to non-decomposable or low probability graphs, giving a poor acceptance ratio. This could possibly be improved by an adaptive sampling scheme, however it is far from clear how this could be efficiently constructed. Furthermore, we could lose the benefits of the local computation of the acceptance ratio.

Due to these difficulties, Jones et al. (2005) and Scott and Carvalho (2008) propose non-MCMC “stochastic search” algorithms for obtaining a representative sample of graphs from the posterior distribution. Although the empirical results of these methods seem promising, their accuracy and theoretical properties remain unknown.

7.2. Directed graphs and dagoids

As noted in section 4, structural Markov laws of ordered directed acyclic graphs are comparatively straightforward to work with, as the parent set of each node can be computed independently. Thus we focus on the dagoid structural Markov law.

Unfortunately, specification of such a MCMC algorithm for dagoids is much more difficult than for the undirected case. Specifically, an individual edge no longer uniquely characterises a neighbouring dagoid, as in Figure 5.

If the directed graph $\vec{\mathcal{G}}^+$ is obtained from the directed graph $\vec{\mathcal{G}}$ by the addition of the edge (u, v) , the only terms in the summation (6.2) that will change are those pertaining to the vertex v , in which case:

$$\begin{aligned} t(\vec{\mathcal{G}}^+) - t(\vec{\mathcal{G}}) &= (\delta_{\{v\} \cup \text{pa}_{\vec{\mathcal{G}}^+}(v)} - \delta_{\text{pa}_{\vec{\mathcal{G}}^+}(v)}) - (\delta_{\{v\} \cup \text{pa}_{\vec{\mathcal{G}}}(v)} - \delta_{\text{pa}_{\vec{\mathcal{G}}}(v)}) \\ &= \delta_{\text{pa}_{\vec{\mathcal{G}}}(v) \cup \{u, v\}} - \delta_{\text{pa}_{\vec{\mathcal{G}}}(v) \cup \{u\}} - \delta_{\text{pa}_{\vec{\mathcal{G}}}(v) \cup \{v\}} + \delta_{\text{pa}_{\vec{\mathcal{G}}}(v)} \end{aligned}$$

In other words, the change in the d-clique vector is determined by the parent set of v in $\vec{\mathcal{G}}$. Therefore, to characterise the neighbouring dagoids (defined as the equivalence classes of the neighbouring graphs), we need to know the parent set of v for each $\vec{\mathcal{G}} \in \mathcal{D}$. Furthermore, as in the undirected case, computing the ratio of probabilities only requires evaluating the parameter on 4 subsets.

Notably, Chickering (2003), Auvray and Wehenkel (2002) and Studený (2005a) develop methods for characterising the neighbouring dagoids. More recently, He, Jia and Yu (2013) developed an MCMC scheme based on this approach. Unfortunately the set of moves is not as easily characterised as in the undirected decomposable case, and the resultant algorithm is considerably more complex.

Another approach is to incorporate an auxiliary variable: this approach was utilised by Madigan et al. (1996), who utilise an auxiliary ordering, and Castelo and Kočka (2004), based on an auxiliary graph. Unfortunately, in both cases the acceptance ratio cannot be computed exactly, so the authors rely on approximations, which may mean that the Markov chain does not have the desired invariant distribution.

References

- ANDERSSON, S. A., MADIGAN, D. and PERLMAN, M. D. (1997a). A characterization of Markov equivalence classes for acyclic digraphs. *Annals of Statistics* **25** 505–541. . [MR1439312 \(99a:62076\)](#)
- ANDERSSON, S. A., MADIGAN, D. and PERLMAN, M. D. (1997b). On the Markov equivalence of chain graphs, undirected graphs, and acyclic digraphs. *Scandinavian Journal of Statistics* **24** 81–102. . [MR1436624 \(98f:62282\)](#)
- ARMSTRONG, H., CARTER, C. K., WONG, K. F. K. and KOHN, R. (2009). Bayesian covariance matrix estimation using a mixture of decomposable graphical models. *Statistics and Computing* **19** 303–316.
- ASMUSSEN, S. and EDWARDS, D. (1983). Collapsibility and response variables in contingency tables. *Biometrika* **70** 567–578. . [MR725370 \(85k:62125\)](#)

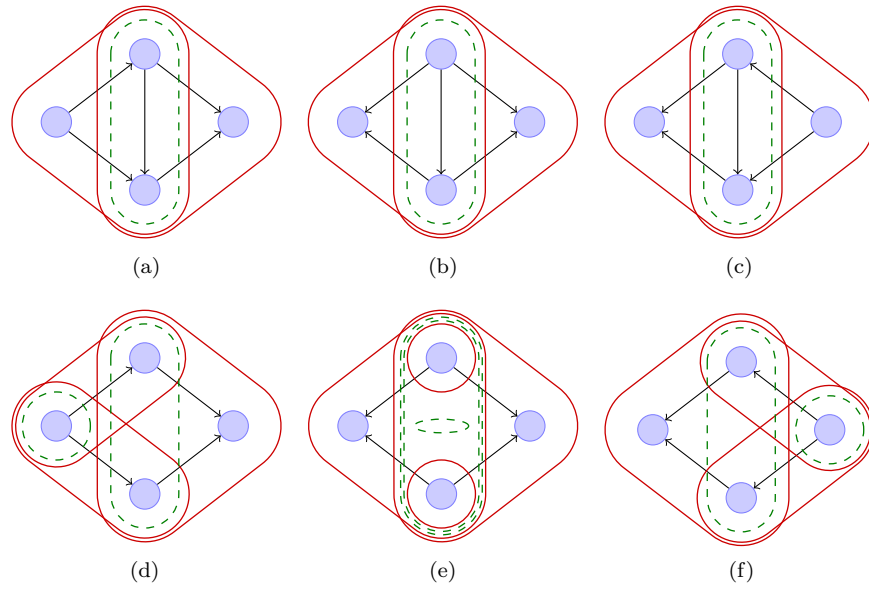


Fig 5: Three Markov equivalent graphs, (a), (b) and (c), in which the same edge removal will result in a transition to a distinct Markov equivalence class, (d), (e) and (f), respectively. The d-cliques (—) and d-separators (---) of each graph are also drawn.

- AUVRAY, V. and WEHENKEL, L. (2002). On the construction of the inclusion boundary neighbourhood for Markov equivalence classes of Bayesian network structures In *Proceedings of the Eighteenth Annual Conference on Uncertainty in Artificial Intelligence* (A. DARWICHE and N. FRIEDMAN, eds.) 26–35. Morgan Kaufmann, San Francisco, CA.
- BORNN, L. and CARON, F. (2011). Bayesian clustering in decomposable graphs. *Bayesian Analysis* **6** 829–845. [MR2869966](#)
- BROOKS, S. P., GIUDICI, P. and ROBERTS, G. O. (2003). Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **65** 3–55. [MR1959092](#) (2004b:65004)
- CASTELO, R. and KOČKA, T. (2004). On inclusion-driven learning of Bayesian networks. *Journal of Machine Learning Research* **4** 527–574. [MR2072261](#)
- CHICKERING, D. M. (1995). A transformational characterization of equivalent Bayesian network structures. In *Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence* (P. Besnard and S. Hanks, eds.) 87–98. Morgan Kaufmann, San Francisco, CA. [MR1615012](#) (99b:68183)
- CHICKERING, D. M. (2003). Optimal structure identification with greedy search. *Journal of Machine Learning Research* **3** 507–554. [MR1991085](#) (2004g:68141)
- DAWID, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)* **41** 1–31. [MR535541](#) (81e:62001)
- DAWID, A. P. (1981). Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika* **68** 265–274. . [MR614963](#) (83m:62083)
- DAWID, A. P. (2001). Separoids: a mathematical framework for conditional independence and irrelevance. *Annals of Mathematics and Artificial Intelligence* **32** 335–372. Representations of uncertainty. . [MR1859870](#) (2002i:62007)
- DAWID, A. P. and LAURITZEN, S. L. (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models. *Annals of Statistics* **21** 1272–1317. . [MR1241267](#) (95c:62015)
- FRYDENBERG, M. (1990). The chain graph Markov property. *Scandinavian Journal of Statistics* **17** 333–353. [MR1096723](#) (92e:60027)
- FRYDENBERG, M. and LAURITZEN, S. L. (1989). Decomposition of maximum likelihood in mixed graphical interaction models. *Biometrika* **76** 539–555. . [MR1040647](#) (91h:62034)
- GIUDICI, P. and GREEN, P. J. (1999). Decomposable graphical Gaussian model determination. *Biometrika* **86** 785–801. . [MR1741977](#) (2001g:62019)
- GREEN, P. J. and THOMAS, A. (2013). Sampling decomposable graphs using a Markov chain on junction trees. *Biometrika* **100** 91–110. . [MR3034326](#)
- HE, Y., JIA, J. and YU, B. (2013). Reversible MCMC on Markov equivalence classes of sparse directed acyclic graphs. *Annals of Statistics* **41** 1742–1779.
- HECKERMAN, D., GEIGER, D. and CHICKERING, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning* **20** 197–243.

- JONES, B., CARVALHO, C., DOBRA, A., HANS, C., CARTER, C. and WEST, M. (2005). Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science* **20** 388–400. . [MR2210226](#)
- KIJIMA, S., KIYOMI, M., OKAMOTO, Y. and UNO, T. (2007). On listing, sampling, and counting the chordal graphs with edge constraints Technical Report, Research Institute for Mathematical Sciences, Kyoto University.
- KIJIMA, S., KIYOMI, M., OKAMOTO, Y. and UNO, T. (2008). On listing, sampling, and counting the chordal graphs with edge constraints. In *Computing and combinatorics. Lecture Notes in Computer Science* **5092** 458–467. Springer, Berlin. . [MR2473446](#)
- LAURITZEN, S. L. (1996). *Graphical models. Oxford Statistical Science Series* **17**. Oxford University Press, New York. [MR1419991](#) (98g:62001)
- LAURITZEN, S. L., SPEED, T. P. and VIJAYAN, K. (1984). Decomposable graphs and hypergraphs. *Journal of the Australian Mathematical Society (Series A)* **36** 12–29. . [MR719998](#) (86g:05075)
- MADIGAN, D. and RAFTERY, A. E. (1994). Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam’s Window. *Journal of the American Statistical Association* **89** 1535–1546.
- MADIGAN, D., ANDERSSON, S. A., PERLMAN, M. D. and VOLINSKY, C. T. (1996). Bayesian model averaging and model selection for markov equivalence classes of acyclic digraphs. *Communications in Statistics. Theory and Methods* **25** 2493–2519.
- MUKHERJEE, S. and SPEED, T. P. (2008). Network inference using informative priors. *Proceedings of the National Academy of Sciences* **105** 14313–14318.
- SCOTT, J. G. and CARVALHO, C. M. (2008). Feature-inclusion stochastic search for Gaussian graphical models. *Journal of Computational and Graphical Statistics* **17** 790–808. . [MR2649067](#)
- STUDENÝ, M. (1997). On marginalization, collapsibility and precollapsibility. In *Distributions with given marginals and moment problems* (V. BENEŠ and J. ŠTĚPÁN, eds.) 191–198. Kluwer Academic Publishers, Dordrecht. [MR1614672](#) (2000g:62127)
- STUDENÝ, M. (2005a). Characterization of inclusion neighbourhood in terms of the essential graph. *International Journal of Approximate Reasoning* **38** 283–309. . [MR2116940](#) (2005h:68150)
- STUDENÝ, M. (2005b). *Probabilistic Conditional Independence Structures*. Springer-Verlag, London.
- STUDENÝ, M. and VOMLEL, J. (2009). A reconstruction algorithm for the essential graph. *International Journal of Approximate Reasoning* **50** 385–413. . [MR2514506](#) (2010e:68159)
- THOMAS, A. and GREEN, P. J. (2009a). Enumerating the decomposable neighbors of a decomposable graph under a simple perturbation scheme. *Computational Statistics & Data Analysis* **53** 1232–1238. . [MR2657086](#)
- THOMAS, A. and GREEN, P. J. (2009b). Enumerating the junction trees of a decomposable graph. *Journal of Computational and Graphical Statistics* **18** 930–940. . [MR2598034](#) (2011d:05183)
- VERMA, T. and PEARL, J. (1990). Equivalence and Synthesis of Causal Models

- In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence* (P. BONISSONE, M. HENRION, L. KANAL and J. LEMMER, eds.) 220–227. Elsevier Science, New York, NY.
- VERMA, T. and PEARL, J. (1992). An Algorithm for Deciding if a Set of Observed Independencies Has a Causal Explanation In *Proceedings of the Eighth Annual Conference on Uncertainty in Artificial Intelligence* (D. DUBOIS, M. WELLMAN, B. D’AMBROSIO and P. SMETS, eds.) 323–330. Morgan Kaufmann, San Mateo, CA.
- WORMALD, N. C. (1985). Counting labelled chordal graphs. *Graphs and Combinatorics* **1** 193–200. . [MR951781](#) ([89e:05109](#))